Partial Solution Manual for "Probabilistic Machine Learning: An Introduction"

Kevin Murphy

September 25, 2021

1 Solutions

Part I
Foundations

2 Solutions

2.1 Conditional independence

2.2 Pairwise independence does not imply mutual independence

We provide two counter examples.

Let X_1 and X_2 be independent binary random variables, and $X_3 = X_1 \oplus X_2$, where $\oplus$ is the XOR operator. We have $p(X_3|X_1, X_2) \neq p(X_3)$, since X_3 can be deterministically calculated from X_1 and X_2. So the variables $\{X_1, X_2, X_3\}$ are not mutually independent. However, we also have $p(X_3|X_1) = p(X_3)$, since without X_2, no information can be provided to X_3. So $X_1 \perp X_3$ and similarly $X_2 \perp X_3$. Hence $\{X_1, X_2, X_3\}$ are pairwise independent.

Here is a different example. Let there be four balls in a bag, numbered 1 to 4. Suppose we draw one at random. Define 3 events as follows:

- X_1: ball 1 or 2 is drawn.
- X_2: ball 2 or 3 is drawn.
- X_3: ball 1 or 3 is drawn.

We have $p(X_1) = p(X_2) = p(X_3) = 0.5$. Also, $p(X_1, X_2) = p(X_2, X_3) = p(X_1, X_3) = 0.25$. Hence $p(X_1, X_2) = p(X_1)p(X_2)$, and similarly for the other pairs. Hence the events are pairwise independent. However, $p(X_1, X_2, X_3) = 0 \neq 1/8 = p(X_1)p(X_2)p(X_3)$.

2.3 Conditional independence iff joint factorizes

2.4 Convolution of two Gaussians is a Gaussian

We follow the derivation of **[Jaynes03]**. Define

$$\phi(x - \mu|\sigma) = \mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma}) \tag{1}$$

where $\phi(z)$ is the pdf of the standard normal. We have

$$p(y) = \mathcal{N}(x_1|\mu_1, \sigma_1^2) \otimes \mathcal{N}(x_2|\mu_2, \sigma_2^2) \tag{2}$$

$$= \int dx_1 \phi(x_1 - \mu_1|\sigma_1)\phi(y - (x_1 - \mu_2)|\sigma_2) \tag{3}$$

Now the product inside the integral can be written as follows

$$\phi(x_1 - \mu_1|\sigma_1)\phi(y - (x_1 - \mu_2)|\sigma_2) = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left\{-\frac{1}{2}\left[\left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2 + \left(\frac{y - x_1 - \mu_2}{\sigma_2}\right)^2\right]\right\} \tag{4}$$

We can bring out the dependency on x_1 by rearranging the quadratic form inside the exponent as follows

$$\left(\frac{x - \mu_1}{\sigma_1}\right)^2 + \left(\frac{y - (x_1 - \mu_2)}{\sigma_2}\right)^2 = (w_1 + w_2)(x_1 - \hat{x}) + \frac{w_1 w_2}{w_1 + w_2}(y - (\mu_1 - \mu_2))^2 \tag{5}$$

where we have defined the precision or weighting terms $w_1 = 1/\sigma_1^2$, $w_2 = 1/\sigma_2^2$, and the term

$$\hat{x} = \frac{w_1\mu_1 + w_2 y - w_2\mu_2}{w_1 + w_2} \tag{6}$$

Note that

$$\frac{w_1 w_2}{w_1 + w_2} = \frac{1}{\sigma_1^2\sigma_2^2}\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2} = \frac{1}{\sigma_1^2 + \sigma_2^2} \tag{7}$$

Hence

$$p(y) = \frac{1}{2\pi\sigma_1\sigma_2} \int \exp\left[-\frac{1}{2}(w_1+w_2)(x_1-\hat{x})^2 - \frac{1}{2}\frac{w_1 w_2}{w_1+w_2}(y-(\mu_1-\mu_2))^2\right] dx_1 \tag{8}$$

The integral over x_1 becomes one over the normalization constant for the Gaussian:

$$\int \exp[-\frac{1}{2}(w_1+w_2)(x_1-\hat{x})^2]\, dx_1 = (2\pi)^{\frac{1}{2}} \left(\frac{\sigma_2^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}\right)^{\frac{1}{2}} \tag{9}$$

Hence

$$p(y) = (2\pi)^{-1}(\sigma_1^2\sigma_2^2)^{-\frac{1}{2}}(2\pi)^{\frac{1}{2}} \left(\frac{\sigma_2^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}\right)^{\frac{1}{2}} \exp[-\frac{1}{2(\sigma_1^2+\sigma_2^2)}(y-\mu_1-\mu_2)^2] \tag{10}$$

$$= (2\pi)^{-\frac{1}{2}} \left(\sigma_1^2+\sigma_2^2\right)^{-\frac{1}{2}} \exp[-\frac{1}{2(\sigma_1^2+\sigma_2^2)}(y-\mu_1-\mu_2)^2] \tag{11}$$

$$= \mathcal{N}(y|\mu_1+\mu_2, \sigma_1^2+\sigma_2^2) \tag{12}$$

2.5 Expected value of the minimum of two rv's

2.6 Variance of a sum

We have

$$\mathbb{V}[X+Y] = E[(X+Y)^2] - (E[X]+E[Y])^2 \tag{13}$$

$$= E[X^2+Y^2+2XY] - (E[X]^2+E[Y]^2+2E[X]E[Y]) \tag{14}$$

$$= E[X^2] - E[X]^2 + E[Y^2] - E[Y]^2 + 2E[XY] - 2E[X]E[Y] \tag{15}$$

$$= \mathbb{V}[X] + \mathbb{V}[Y] + 2\mathrm{Cov}[X,Y] \tag{16}$$

If X and Y are independent, then $\mathrm{Cov}[X,Y] = 0$, so $\mathbb{V}[X+Y] = \mathbb{V}[X] + \mathbb{V}[Y]$.

2.7 Deriving the inverse gamma density

2.8 Mean, mode, variance for the beta distribution

For the mode we can use simple calculus, as follows. Let $f(x) = \mathrm{Beta}(x|a,b)$. We have

$$0 = \frac{df}{dx} \tag{17}$$

$$= (a-1)x^{a-2}(1-x)^{b-1} - (b-1)x^{a-1}(1-x)^{b-2} \tag{18}$$

$$= (a-1)(1-x) - (b-1)x \tag{19}$$

$$x = \frac{a-1}{a+b-2} \tag{20}$$

For the mean we have

$$\mathbb{E}[\theta|\mathcal{D}] = \int_0^1 \theta p(\theta|\mathcal{D}) d\theta \tag{21}$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int \theta^{(a+1)-1}(1-\theta)^{b-1} d\theta \tag{22}$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+1+b)} = \frac{a}{a+b} \tag{23}$$

where we used the definition of the Gamma function and the fact that $\Gamma(x+1) = x\Gamma(x)$.

We can find the variance in the same way, by first showing that

$$\mathbb{E}\left[\theta^2\right] = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \frac{\Gamma(a+2+b)}{\Gamma(a+2)\Gamma(b)} \theta^{(a+2)-1}(1-\theta)^{b-1} d\theta \tag{24}$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+2+b)}{\Gamma(a+2)\Gamma(b)} = \frac{a}{a+b} \frac{a+1}{a+1+b} \tag{25}$$

Now we use $\mathbb{V}[\theta] = \mathbb{E}\left[\theta^2\right] - \mathbb{E}[\theta]^2$ and $\mathbb{E}[\theta] = a/(a+b)$ to get the variance.

2.9 Bayes rule for medical diagnosis

2.10 Legal reasoning

Let E be the evidence (the observed blood type), and I be the event that the defendant is innocent, and $G = \neg I$ be the event that the defendant is guilty.

1. The prosecutor is confusing $p(E|I)$ with $p(I|E)$. We are told that $p(E|I) = 0.01$ but the relevant quantity is $p(I|E)$. By Bayes rule, this is

$$p(I|E) = \frac{p(E|I)p(I)}{p(E|I)p(I) + p(E|G)p(G)} = \frac{0.01p(I)}{0.01p(I) + (1-p(I))} \tag{26}$$

   since $p(E|G) = 1$ and $p(G) = 1 - p(I)$. So we cannot determine $p(I|E)$ without knowing the prior probability $p(I)$. So $p(E|I) = p(I|E)$ only if $p(G) = p(I) = 0.5$, which is hardly a presumption of innocence.

   To understand this more intuitively, consider the following isomorphic problem (from http://en.wikipedia.org/wiki/Prosecutor's_fallacy):

   > A big bowl is filled with a large but unknown number of balls. Some of the balls are made of wood, and some of them are made of plastic. Of the wooden balls, 100 are white; out of the plastic balls, 99 are red and only 1 are white. A ball is pulled out at random, and observed to be white.

   Without knowledge of the relative proportions of wooden and plastic balls, we cannot tell how likely it is that the ball is wooden. If the number of plastic balls is far larger than the number of wooden balls, for instance, then a white ball pulled from the bowl at random is far more likely to be a white plastic ball than a white wooden ball — even though white plastic balls are a minority of the whole set of plastic balls.

2. The defender is quoting $p(G|E)$ while ignoring $p(G)$. The prior odds are

$$\frac{p(G)}{p(I)} = \frac{1}{799,999} \tag{27}$$

   The posterior odds are

$$\frac{p(G|E)}{p(I|E)} = \frac{1}{7999} \tag{28}$$

   So the evidence has increased the odds of guilt by a factor of 1000. This is clearly relevant, although perhaps still not enough to find the suspect guilty.

2.11 Probabilities are sensitive to the form of the question that was used to generate the answer

2.12 Normalization constant for a 1D Gaussian

Following the first hint we have

$$Z^2 = \int_0^{2\pi} \int_0^{\infty} r \exp\left(-\frac{r^2}{2\sigma^2}\right) dr d\theta \tag{29}$$

$$= \left[\int_0^{2\pi} d\theta\right] \left[\int_0^{\infty} r \exp\left(-\frac{r^2}{2\sigma^2}\right) dr\right] \tag{30}$$

$$= (2\pi)I \tag{31}$$

where I is the inner integral

$$I = \int_0^{\infty} r \exp\left(-\frac{r^2}{2\sigma^2}\right) \tag{32}$$

Following the second hint we have

$$I = -\sigma^2 \int -\frac{r}{\sigma^2} e^{-r^2/2\sigma^2} dr \tag{33}$$

$$= -\sigma^2 \left[e^{-r^2/2\sigma^2}\right]_0^{\infty} \tag{34}$$

$$= -\sigma^2[0-1] = \sigma^2 \tag{35}$$

Hence

$$Z^2 = 2\pi\sigma^2 \tag{36}$$

$$Z = \sigma\sqrt{(2\pi)} \tag{37}$$

3 Solutions

3.1 Uncorrelated does not imply independent

3.2 Correlation coefficient is between -1 and +1

We have

$$0 \leq \mathbb{V}\left[\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right] \tag{38}$$

$$= \mathbb{V}\left[\frac{X}{\sigma_X}\right] + \mathbb{V}\left[\frac{Y}{\sigma_Y}\right] + 2\text{Cov}\left[\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}\right] \tag{39}$$

$$= \frac{\mathbb{V}[X]}{\sigma_X^2} + \frac{\mathbb{V}[Y]}{\sigma_Y^2} + 2\text{Cov}\left[\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}\right] \tag{40}$$

$$= 1 + 1 + 2\rho = 2(1+\rho) \tag{41}$$

Hence $\rho \geq -1$. Similarly,

$$0 \leq \mathbb{V}\left[\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}\right] = 2(1-\rho) \tag{42}$$

so $\rho \leq 1$.

3.3 Correlation coefficient for linearly related variables is ± 1

3.4 Linear combinations of random variables

1. Let $\boldsymbol{y} = \mathbf{A}\boldsymbol{x}$. Then

$$\text{Cov}[y] = \mathbb{E}\left[(\boldsymbol{y} - \mathbb{E}[\boldsymbol{y}])(\boldsymbol{y} - \mathbb{E}[\boldsymbol{y}])^T\right] \tag{43}$$

$$= \mathbb{E}\left[(\mathbf{A}\boldsymbol{x} - \mathbf{A}\boldsymbol{m})(\mathbf{A}\boldsymbol{x} - \mathbf{A}\boldsymbol{m})^T\right] \tag{44}$$

$$= \mathbf{A}\mathbb{E}\left[(\boldsymbol{x} - \boldsymbol{m})(\boldsymbol{x} - \boldsymbol{m})^T\right]\mathbf{A}^T \tag{45}$$

$$= \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T \tag{46}$$

2. $\mathbf{C} = \mathbf{AB}$ has entries $c_{ij} = \sum_k a_{ik}b_{kj}$. The diagonal elements of C (when $i = j$) are given by $\sum_k a_{ik}b_{ki}$. So the sum of the diagonal elements is $\text{tr}(\mathbf{AB}) = \sum_{ik} a_{ik}b_{ki}$ which is symmetric in $\mathbf{A}$ and $\mathbf{B}$.

3. We have

$$\mathbb{E}\left[\boldsymbol{x}^T\mathbf{A}\boldsymbol{x}\right] = \mathbb{E}\left[\text{tr}(\boldsymbol{x}^T\mathbf{A}\boldsymbol{x})\right] = \mathbb{E}\left[\text{tr}(\mathbf{A}\boldsymbol{x}\boldsymbol{x}^T)\right] \tag{47}$$

$$= \text{tr}(\mathbf{A}\mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^T\right]) = \text{tr}(\mathbf{A}(\boldsymbol{\Sigma} + \boldsymbol{m}\boldsymbol{m}^T)) \tag{48}$$

$$= \text{tr}(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{m}^T\mathbf{A}\boldsymbol{m} \tag{49}$$

3.5 Gaussian vs jointly Gaussian

1. For the mean, we have

$$\mathbb{E}[Y] = \mathbb{E}[WX] = \mathbb{E}[X]\,\mathbb{E}[X] = 0 \tag{50}$$

For the variance, we have

$$\mathbb{V}[Y] = \mathbb{E}[\mathbb{V}[Y|W]] + \mathbb{V}[\mathbb{E}[Y|W]] \tag{51}$$

$$= \mathbb{E}[W[\mathbb{V}[X]]W] + \mathbb{V}[W\mathbb{E}[X]] \tag{52}$$

$$= \mathbb{E}\left[W^2\right] + 0 = 1 \tag{53}$$

To show it's Gaussian, we note that Y is a linear combination of Gaussian rv's.

2. To show that $\text{Cov}[X, Y] = 0$, we use the rule of iterated expectation. First we have

$$\mathbb{E}[XY] = \mathbb{E}[\mathbb{E}[XY|W]] \tag{54}$$
$$= \sum_{w \in \{-1,1\}} p(w)\mathbb{E}[XY|w] \tag{55}$$
$$= -1 \cdot 0.5 \cdot \mathbb{E}[X \cdot -X] + 1 \cdot 0.5 \cdot \mathbb{E}[X \cdot X] \tag{56}$$
$$== 0 \tag{57}$$

Then we have

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|W]] \tag{58}$$
$$= \sum_{w \in \{-1,1\}} p(w)\mathbb{E}[Y|w] \tag{59}$$
$$= 0.5 \cdot \mathbb{E}[-X] + 0.5 \cdot \mathbb{E}[X] \tag{60}$$
$$= 0 \tag{61}$$

Hence

$$\text{Cov}[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0 \tag{62}$$

So X and Y are uncorrelated even though they are dependent.

3.6 Normalization constant for a multidimensional Gaussian

Let $\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$, so

$$\boldsymbol{\Sigma}^{-1} = \mathbf{U}^{-T}\boldsymbol{\Lambda}^{-1}\mathbf{U}^{-1} = \mathbf{U}\boldsymbol{\Lambda}^{-1}\mathbf{U} = \sum_{i=1}^{p} \frac{1}{\lambda_i}\boldsymbol{u}_i\boldsymbol{u}_i^T \tag{63}$$

Hence

$$(\boldsymbol{x} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \left(\sum_{i=1}^{p} \frac{1}{\lambda_i}\boldsymbol{u}_i\boldsymbol{u}_i^T\right)(\boldsymbol{x} - \boldsymbol{\mu}) \tag{64}$$
$$= \sum_{i=1}^{p} \frac{1}{\lambda_i}(\boldsymbol{x} - \boldsymbol{\mu})^T\boldsymbol{u}_i\boldsymbol{u}_i^T(\boldsymbol{x} - \boldsymbol{\mu}) = \sum_{i=1}^{p} \frac{y_i^2}{\lambda_i} \tag{65}$$

where $y_i \triangleq \boldsymbol{u}_i^T(\boldsymbol{x} - \boldsymbol{\mu})$. The $\boldsymbol{y}$ variables define a new coordinate system that is shifted (by $\boldsymbol{\mu}$) and rotated (by $\mathbf{U}$) with respect to the original x coordinates: $\boldsymbol{y} = \mathbf{U}(\boldsymbol{x} - \boldsymbol{\mu})$. Hence $\boldsymbol{x} = \mathbf{U}^T\boldsymbol{y} + \boldsymbol{\mu}$.

The Jacobian of this transformation, from $\boldsymbol{y}$ to $\boldsymbol{x}$, is a matrix with elements

$$J_{ij} = \frac{\partial x_i}{\partial y_j} = U_{ji} \tag{66}$$

so $\mathbf{J} = \mathbf{U}^T$ and $|\mathbf{J}| = 1$.

So

$$\int \exp(-\frac{1}{2}(\boldsymbol{x} - \mu)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}))d\boldsymbol{x} = \int \prod_i \exp(-\frac{1}{2}\sum_i \frac{y_i^2}{\lambda_i})dy_i|\mathbf{J}| \tag{67}$$
$$= \prod_i \sqrt{2\pi\lambda_i} = |2\pi\boldsymbol{\Sigma}| \tag{68}$$

3.7 Sensor fusion with known variances in 1d

Define the sufficient statistics as

$$\overline{y}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} y_i^{(1)}, \;\; \overline{y}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} y_i^{(2)}, \tag{69}$$

Define the prior as

$$\mu_\mu = 0, \Sigma_\mu = \infty \tag{70}$$

Define the likelihood as

$$\mathbf{A} = \begin{pmatrix}1\\1\end{pmatrix}, \boldsymbol{\mu}_y = \begin{pmatrix}0\\0\end{pmatrix}, \boldsymbol{\Sigma}_y = \begin{pmatrix}\frac{v_1}{n_1} & 0\\ 0 & \frac{v_2}{n_2}\end{pmatrix} \tag{71}$$

Now we just apply the equations. The posterior precision is a sum of the precisions of each sensor:

$$\Sigma_{\mu|y}^{-1} = \mathbf{A}^T\boldsymbol{\Sigma}_y^{-1}\mathbf{A} = \begin{pmatrix}1 & 1\end{pmatrix}\begin{pmatrix}\frac{v_1}{n_1} & 0\\ 0 & \frac{v_2}{n_2}\end{pmatrix}\begin{pmatrix}1\\1\end{pmatrix} = \frac{n_1}{v_1} + \frac{n_2}{v_2} \tag{72}$$

The posterior mean is a weighted sum of the observed values from each sensor:

$$\mu_{\mu|y} = \Sigma_{\mu|y}^{-1}\left(\begin{pmatrix}1 & 1\end{pmatrix}\begin{pmatrix}\frac{v_1}{n_1} & 0\\ 0 & \frac{v_2}{n_2}\end{pmatrix}\begin{pmatrix}\overline{y}_1\\ \overline{y}_2\end{pmatrix}\right) = \Sigma_{\mu|y}^{-1}\left(\frac{n_1\overline{y}_1}{v_1} + \frac{n_2\overline{y}_2}{v_2}\right) \tag{73}$$

3.8 Show that the Student distribution can be written as a Gaussian scale mixture

$$p(w|\mu,a,b) = \int_0^\infty \mathcal{N}(w|\mu,\alpha^{-1})\mathrm{Ga}(\alpha|a,b)d\alpha \tag{74}$$

$$= \int_0^\infty \frac{b^a e^{-b\alpha}\alpha^{a-1}}{\Gamma(a)}\left(\frac{\alpha}{2\pi}\right)^{\frac{1}{2}}\exp\left(-\frac{\alpha}{2}(w-\mu)^2\right)d\alpha \tag{75}$$

$$= \frac{b^a}{\Gamma(a)}\left(\frac{1}{2\pi}\right)^{\frac{1}{2}}\int_0^\infty \alpha^{a-\frac{1}{2}}\exp\left[-\alpha\left(b+\frac{1}{2}(w-\mu)^2\right)\right]d\alpha \tag{76}$$

Let us define $\Delta = b + (w-\mu)^2/2$ and $z = \alpha\Delta$. Then, using $dz = \Delta d\alpha$, and the definition of the Gamma function, $\int_0^\infty u^{x-1}e^{-u}du = \Gamma(x)$, the integral becomes

$$\int_0^\infty \left(\frac{z}{\Delta}\right)^{a+\frac{1}{2}-1} e^{-z}\Delta^{-1}dz = \Delta^{-a-\frac{1}{2}}\Gamma(a+\frac{1}{2}) \tag{77}$$

so we have

$$p(w|\mu,a,b) = \frac{\Gamma(a+1/2)}{\Gamma(a)}b^a\left(\frac{1}{2\pi}\right)^{\frac{1}{2}}\Delta^{-a-\frac{1}{2}} \tag{78}$$

Let us define $a = \nu/2$ and $b = \nu/(2\lambda)$. Then we have

$$p(w|\mu,a,b) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)}\left(\frac{\nu}{2\lambda}\right)^{\nu/2}\left(\frac{1}{2\pi}\right)^{\frac{1}{2}}\left(\frac{\nu}{2\lambda}+\frac{(w-\mu)^2}{2}\right)^{-(\nu+1)/2} \tag{79}$$

$$= \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)}\left(\frac{\nu}{2\lambda}\right)^{\nu/2}\left(\frac{1}{2\pi}\right)^{\frac{1}{2}}\left(\frac{\nu}{2\lambda}\left[1+\frac{\lambda}{\nu}(w-\mu)^2\right]\right)^{-(\nu+1)/2} \tag{80}$$

$$= \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)}\left(\frac{\lambda}{\nu\pi}\right)^{\frac{1}{2}}\left[1+\frac{\lambda}{\nu}(w-\mu)^2\right]^{-(\nu+1)/2} \tag{81}$$

$$= \mathcal{T}_\nu(w|\mu,\lambda^{-1}) \tag{82}$$

4 Solutions

4.1 MLE for the univariate Gaussian

4.2 MAP estimation for 1D Gaussians

4.3 Gaussian posterior credible interval

We want an interval that satisfies

$$p(\ell \leq \mu_n \leq u|D) \geq 0.95 \tag{83}$$

where

$$\ell = \mu_n + \Phi^{-1}(0.025)\sigma_n = \mu_n - 1.96\sigma_n \tag{84}$$

$$u = \mu_n + \Phi^{-1}(0.9755)\sigma_n = \mu_n + 1.96\sigma_n \tag{85}$$

where Φ is the cumulative distribution function for the standard normal $\mathcal{N}(0,1)$ distribution, and $\Phi^{-1}(0.025)$ is the value below which 2.5% of the probability mass lies (in Matlab, `norminv(0.025)=-1.96`). and $\Phi^{-1}(0.975)$ is the value below which 97.5% of the probability mass lies (in Matlab, `norminv(0.975)=1.96`). We want to find n such that

$$u - \ell = 1 \tag{86}$$

Hence we solve

$$2(1.96)\sigma_n = 1 \tag{87}$$

$$\sigma_n^2 = \frac{1}{4(1.96)^2} \tag{88}$$

where

$$\sigma_n^2 = \frac{\sigma^2\sigma_0^2}{n\sigma_0^2 + \sigma^2} \tag{89}$$

Hence

$$n\sigma_0^2 + \sigma^2 = (\sigma^2\sigma_0^2)4(1.96)^2 \tag{90}$$

$$n = \frac{\sigma^2(\sigma_0^2 4(1.96)^2 - 1)}{\sigma_0^2} \tag{91}$$

$$= \frac{4(9 \times (1.96)^2 - 1)}{9} = 61.0212 \tag{92}$$

Hence we need at least $n \geq 62$ samples.

4.4 BIC for Gaussians

4.5 BIC for a 2d discrete distribution

1. The joint distribution is $p(x, y|\boldsymbol{\theta}) = p(x|\theta_1)p(y|x, \theta_2)$:

	$y=0$	$y=1$
$x=0$	$(1-\theta_1)\theta_2$	$(1-\theta_1)(1-\theta_2)$
$x=1$	$\theta_1(1-\theta_2)$	$\theta_1\theta_2$

2. The log likelihood is

$$\log p(\mathcal{D}|\boldsymbol{\theta}) = \sum_i \log p(x_i|\theta_1) + \sum_i \log p(y_i|x_i, \theta_2) \tag{93}$$

Hence we can optimize each term separately. For θ_1, we have

$$\hat{\theta}_1 = \frac{\sum_i I(x_i = 1)}{n} = \frac{N(x=1)}{N} = \frac{4}{7} = 0.5714 \tag{94}$$

For θ_2, we have

$$\hat{\theta}_2 = \frac{\sum_i I(x_i = y_i)}{n} = \frac{N(x=y)}{N} = \frac{4}{7} \tag{95}$$

The likelihood is

$$p(\mathcal{D}|\hat{\boldsymbol{\theta}}, M_2) = (\frac{4}{7})^{N(x=1)}(\frac{3}{7})^{N(x=0)}(\frac{4}{7})^{N(x=y)}(\frac{3}{7})^{N(x\neq y)} \tag{96}$$

$$= (\frac{4}{7})^4(\frac{3}{7})^3(\frac{4}{7})^4(\frac{3}{7})^3 \tag{97}$$

$$= (\frac{4}{7})^8(\frac{3}{7})^6 \approx 7.04 \times 10^{-5} \tag{98}$$

3. The table of joint counts is

	$y=0$	$y=1$
$x=0$	2	1
$x=1$	2	2

We can think of this as a multinomial distribution with 4 states. Normalizing the counts gives the MLE:

	$y=0$	$y=1$
$x=0$	2/7	1/7
$x=1$	2/7	2/7

The likelihood is

$$p(\mathcal{D}|\hat{\boldsymbol{\theta}}, M_4) = \theta_{00}^{N(x=0,y=0)}\theta_{01}^{N(x=0,y=1)}\theta_{10}^{N(x=1,y=0)}\theta_{11}^{N(x=1,y=1)} = (\frac{2}{7})^2(\frac{1}{7})^1(\frac{2}{7})^2(\frac{2}{7})^2 \tag{99}$$

$$= (\frac{2}{7})^6(\frac{1}{7})^1 \approx 7.77 \times 10^{-5} \tag{100}$$

Thus is higher than the previous likelihood, because the model has more parameters.

4. For M_4, when we omit case 7, we will have $\hat{\theta}_{01} = 0$, so $p(x_7, y_7|m_4, \hat{\theta}) = 0$, so $L(m_4) = -\infty$. However, $L(m_2)$ will be finite, since all counts remain non zero when we leave out a single case. Hence CV will prefer M_2, since M_4 is overfitting.

5. The BIC score is

$$BIC(m) = \log p(\mathcal{D}|\hat{\boldsymbol{\theta}}, m) - \frac{\text{dof}(m)}{2}\log n \tag{101}$$

where $n = 7$. For M_2, we have dof $= 2$, so

$$BIC(m_2) = 8\log(\frac{4}{7}) + 6\log(\frac{3}{7}) - \frac{2}{2}\log 7 = -11.5066 \tag{102}$$

For M_4, we have dof $= 3$ because of the sum-to-one constraint, so

$$BIC(m_4) = 6\log(\frac{2}{7}) + 1\log(\frac{1}{7}) - \frac{3}{2}\log 7 = -12.3814 \tag{103}$$

So BIC also prefers m_2.

4.6 A mixture of conjugate priors is conjugate

4.7 ML estimator σ^2_{mle} is biased

Because the variance of any random variable R is given by $var(R) = E[R^2] - (E[R])^2$, the expected value of the square of a Gaussian random variable X_i with mean μ and variance σ^2 is $E[X_i^2] = var(X_i) + (E[X_i])^2 = \sigma^2 + \mu^2$.

$$\begin{aligned}
E_{X_1,\ldots,X_n \sim \mathcal{N}(\mu,\sigma)}[\sigma^2(X_1,\ldots,X_n)] &= E[\frac{1}{n}\sum_{i=1}^{n}(X_i - \frac{\sum_{j=1}^{n} X_j}{n})^2] \\
&= \frac{1}{n}\sum_{i=1} nE[(X_i - \frac{\sum_{j=1}^{n} X_j}{n})^2] \\
&= \frac{1}{n}\sum_{i=1} nE[(X_i - \frac{\sum_{j=1}^{n} X_j}{n})(X_i - \frac{\sum_{j=1}^{n} X_j}{n})] \\
&= \frac{1}{n}\sum_{i=1} nE[X_i^2 - \frac{2}{n}X_i\sum_{j=1}^{n} X_j + \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n} X_j X_k] \\
&= \frac{1}{n}\sum_{i=1} nE[X_i^2] - \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} E[X_i X_j] + \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}[X_j X_k]
\end{aligned}$$

Consider the two summations $\sum_{i=1}^{n}\sum_{j=1}^{n} E[X_i X_j]$ and $\sum_{j=1}^{n}\sum_{k=1}^{n}[X_j X_k]$. Of the n^2 terms in each of these summations, n of them satisfy $i = j$ or $j = k$, so these terms are of the form $E[X_i^2]$. By linearity of expectation, these terms contribute $nE[X_i^2]$ to the sum. The remaining $n^2 - n$ terms are of the form $E[X_i X_j]$ or $E[X_j X_k]$ for $i \neq j$ or $j \neq k$. Because the X_i are independent samples, it follows from linearity of expectation that these terms contribute $(n^2 - n)E[X_i]E[X_j]$ to the summation.

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{n} E[X_i X_j] &= \sum_{j=1}^{n}\sum_{k=1}^{n} E[X_j X_k] \\
&= nE[X_i^2] + (n^2 - n)E[X_i][X_j] \\
&= n(\sigma^2 + \mu^2) + (n^2 - n)\mu\mu = n\sigma^2 + n\mu^2 + n^2\mu^2 - n\mu^2 \\
&= n\sigma^2 + n^2\mu^2
\end{aligned}$$

$$\begin{aligned}
&E_{X_1,\ldots,X_n \sim \mathcal{N}(\mu,\sigma)}[\sigma^2(X_1,\ldots,X_n)] = \\
&\frac{1}{n}\sum_{i} = 1^n(\sigma^2 + \mu^2) - \frac{2}{n^2}(n\sigma^2 + n^2\mu^2) + \frac{1}{n^3}\sum_{i=1}^{n}(n\sigma^2 + n^2\mu^2) \\
&= \frac{1}{n}(n\sigma^2 + n\mu^2) - 2\frac{\sigma^2}{n} - 2\mu^2 + \frac{1}{n^3}(n^2\sigma^2 + N^3\mu^2) \\
&= \sigma^2 + \mu^2 - 2\frac{\sigma^2}{n} - 2\mu^2 + \frac{\sigma^2}{n} + \mu^2 \\
&= \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2
\end{aligned}$$

Since the expected value of $\hat{\sigma^2}(X_1,\ldots,X_n)$ is not equal to the actual variance σ^2, σ^2 is not an unbiased estimator. In fact, the maximum likelihood estimator tends to underestimate the variance. This is not surprising: consider the case of only a single sample: we will never detect any variance. If there are multiple samples, we will detect variance, but since our estimate for the mean will tend to be shifted from the true mean in the direction of our samples, we will tend to underestimate the variance.

4.8 Estimation of σ^2 when μ is known

4.9 Variance and MSE of the unbiased estimator for Gaussian variance

5 Solutions

5.1 Reject option in classifiers

1. We have to choose between rejecting, with risk λ_r, and choosing the most probable class, $j_{max} = \arg\max_j p(Y = j|\boldsymbol{x})$, which has risk

$$\lambda_s \sum_{j \neq j_{max}} p(Y = j|\boldsymbol{x}) = \lambda_s(1 - p(Y = j_{max}|\boldsymbol{x})) \tag{104}$$

Hence we should pick j_{max} if

$$\lambda_r \geq \lambda_s(1 - p(Y = j_{max}|\boldsymbol{x})) \tag{105}$$

$$\frac{\lambda_r}{\lambda_s} \geq (1 - p(Y = j_{max}|\boldsymbol{x})) \tag{106}$$

$$p(Y = j_{max}|\boldsymbol{x}) \geq 1 - \frac{\lambda_r}{\lambda_s} \tag{107}$$

otherwise we should reject.

For completeness, we should prove that when we decide to choose a class (and not reject), we always pick the most probable one. If we choose a non-maximal category $k \neq j_{max}$, the risk is

$$\lambda_s \sum_{j \neq k} p(Y = j|\boldsymbol{x}) = \lambda_s(1 - p(Y = k|\boldsymbol{x})) \geq \lambda_s(1 - p(Y = j_{max}|\boldsymbol{x})) \tag{108}$$

which is always bigger than picking j_{max}.

2. If $\lambda_r/\lambda_s = 0$, there is no cost to rejecting, so we always reject. As $\lambda_r/\lambda_s \to 1$, the cost of rejecting increases. We find $p(Y = j_{max}|\boldsymbol{x}) \geq 1 - \frac{\lambda_r}{\lambda_s}$ is always satisfied, so we always accept the most probable class.

5.2 Newsvendor problem

5.3 Bayes factors and ROC curves

5.4 Posterior median is optimal estimate under L1 loss

To prove this, we expand the posterior expected loss as follows:

$$\rho(a|\boldsymbol{x}) = E_{\theta|x}|\theta - a| = \int_{\theta \geq a} (\theta - a)p(\theta|x)d\theta + \int_{\theta \leq a} (a - \theta)p(\theta|x)d\theta \tag{109}$$

$$= \int_a^{\infty} (\theta - a)p(\theta|x)d\theta + \int_{-\infty}^{a} (a - \theta)p(\theta|x)d\theta \tag{110}$$

Now recall the rule to differentiate under the integral sign:

$$\frac{d}{da}\int_{A(a)}^{B(a)} \phi(a, \theta)d\theta = \int_{A(a)}^{B(a)} \phi'(a, \theta)d\theta + \phi(a, B(a))B'(a) + \phi(a, A(a))A'(a) \tag{111}$$

where $\phi'(a, \theta) = \frac{d}{da}\phi(a, \theta)$. Applying this to the first integral in Equation 110, with $A(a) = a$, $B(a) = \infty$, $\phi(a, \theta) = (\theta - a)p(\theta|x)dy$, we have

$$\int_a^{\infty} (\theta - a)p(\theta|x)d\theta = \int_a^{\infty} -p(\theta|x)d\theta + 0 + 0 \tag{112}$$

Analogously, one can show

$$\int_{-\infty}^{a} (a-\theta)p(\theta|x)d\theta = \int_{-\infty}^{a} p(\theta|x)d\theta \tag{113}$$

Hence

$$\rho'(a|\boldsymbol{x}) = -\int_{a}^{\infty} p(\theta|x)d\theta + \int_{-\infty}^{a} p(\theta|x)d\theta \tag{114}$$

$$= -P(\theta \geq a|x) + P(\theta \leq a|x) = 0 \tag{115}$$

So the value of a that makes $\rho'(a|\boldsymbol{x}) = 0$ satisfies

$$P(\theta \geq a|\boldsymbol{x}) = P(\theta \leq a|\boldsymbol{x}) \tag{116}$$

Hence the optimal a is the posterior median.

6 Solutions

6.1 Expressing mutual information in terms of entropies

6.2 Relationship between $D(p||q)$ and χ^2 statistic

We have

$$D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \tag{117}$$

$$= \sum_x (Q(x) + \Delta(x)) \log \left(1 + \frac{\Delta(x)}{q(x)}\right) \tag{118}$$

$$= \sum_x (Q(x) + \Delta(x)) \left(\frac{\Delta(x)}{q(x)} - \frac{\Delta(x)^2}{2q(x)} + \cdots\right) \tag{119}$$

$$= \sum_x \Delta(x) + \frac{\Delta(x)^2}{q(x)} - \frac{\Delta(x)^2}{2q(x)} + \cdots \tag{120}$$

$$= 0 + \sum_x \frac{\Delta(x)^2}{2q(x)} + \cdots \tag{121}$$

since

$$\sum_x \Delta(x) = \sum_x p(x) - q(x) = 0 \tag{122}$$

6.3 Fun with entropies

6.4 Forwards vs reverse KL divergence

1. We have

$$\mathbb{KL}(p\|q) = \sum_{xy} p(x,y)[\log p(x,y) - \log q(x) - \log q(y)] \tag{123}$$

$$= \sum_{xy} p(x,y) \log p(x,y) - \sum_x p(x) \log q(x) - \sum_y p(y) \log q(y) \tag{124}$$

We can optimize wrt $q(x)$ and $q(y)$ separately. Imposing a Lagrange multiplier to enforce the constraint that $\sum_x q(x) = 1$ we have the Lagrangian

$$\mathcal{L}(q, \lambda) = \sum_x p(x) \log q(x) + \lambda(1 - \sum_x q(x)) \tag{125}$$

Taking derivatives wrt $q(x)$ (thinking of the function as a finite length vector, for simplicity), we have

$$\frac{\partial \mathcal{L}}{\partial q(x)} = \frac{p(x)}{q(x)} - \lambda = 0 \tag{126}$$

$$q(x) = \frac{p(x)}{\lambda} \tag{127}$$

Summing both sides over x we get $\lambda = 1$ and hence

$$q(x) = p(x) \tag{128}$$

Analogously, $q(y) = p(y)$.

2. We require $q(x,y) = 0$ whenever $p(x,y) = 0$, otherwise $\log q(x,y)/p(x,y) = \infty$. Since $q(x,y) = q_x(x)q_y(y)$, it must be that $q_x(x) = q_y(y)$ whenever $x = y$, and hence $q_x = q_y$ are the same distribution. There are only 3 possible distributions that put 0s in the right places and yet sum to 1. The first is:

		x				
		1	2	3	4	q(y)
	1	0	0	0	0	0
y	2	0	0	0	0	0
	3	0	0	1	0	1
	4	0	0	0	0	0
	q(x)	0	0	1	0	

The second one is

		x				
		1	2	3	4	q(y)
	1	0	0	0	0	0
y	2	0	0	0	0	0
	3	0	0	0	0	0
	4	0	0	0	1	1
	q(x)	0	0	0	1	

For both of these, we have $\mathbb{KL}(q\|p) = 1 \times \log \frac{1}{1/4} = \log 4$. Furthermore, any slight perturbation of these probabilities away from the designated values will cause the KL to blow up, meaning these are local minima.

The final local optimum is

		x				
		1	2	3	4	q(y)
	1	1/4	1/4	0	0	1/2
y	2	1/4	1/4	0	0	1/2
	3	0	0	0	0	0
	4	0	0	0	0	0
	q(x)	1/2	1/2	0	0	

This has $\mathbb{KL}(q\|p) = 4(\frac{1}{4}\log\frac{1/4}{1/8}) = \log 2$, so this is actually the global optimum.

To see that there are no other solutions, one can do a case analysis, and see that any other distribution will not put 0s in the right places. For example, consider this:

		x				
		1	2	3	4	q(y)
	1	1/4	0	1/4	0	1/2
y	2	0	0	0	0	0
	3	1/4	0	1/4	0	1/2
	4	0	0	0	0	0
	q(x)	1/2	0	1/2	0	

Obviously if we set $q(x,y) = p(x)p(y) = 1/16$, we get $\mathbb{KL}(q\|p) = \infty$.

7 Solutions

7.1 Orthogonal matrices

1. Let $c = \cos(\alpha)$ and $s = \sin(\alpha)$. Using the fact that $c^2 + s^2 = 1$, we have

$$\mathbf{R}^T\mathbf{R} = \begin{pmatrix} c & s & 0 \\ -s & c & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} c & -s & 0 \\ s & c & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} c^2+s^2+0 & -cs+sc+0 & 0 \\ -sc+sc+0 & c^2+s^2+0 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{129}$$

2. The z-axis $\boldsymbol{v} = (0, 0, 1)$ is not affected by a rotation around z. We can easily check that $(0, 0, 1)$ is an eigenvector with eigenvalue 1 as follows:

$$\begin{pmatrix} c & s & 0 \\ -s & c & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \tag{130}$$

Of course, $(0, 0, -1)$ is also a valid solution. We can check this using the symbolic math toolbox in Matlab:

```
syms c s x y z
S=solve('c*x-s*y=x','s*x+c*y=y','x^2+y^2+z^2=1')
>> S.x
ans =
 0
 0
>> S.y
ans =
 0
 0
>> S.z
ans =
  1
 -1
```

If we ignore the unit norm constraint, we find that $(0, 0, z)$ is a solution for any $z \in \mathbb{R}$. We can see this as follows:

$$\begin{pmatrix} c & s & 0 \\ -s & c & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} \tag{131}$$

becomes

$$cx - sy = x \tag{132}$$
$$sx + cy = y \tag{133}$$
$$z = z \tag{134}$$

Solving gives

$$y = \frac{x(c-1)}{s} \tag{135}$$
$$sx + cy = sx + \frac{cx(x-1)}{s} \tag{136}$$
$$y = \frac{x(c-1)}{s} \tag{137}$$

and hence $x = y = 0$. We can check this in Matlab:

```
 S=solve('c*x-s*y=x','s*x+c*y=y')
S =
    x: [1x1 sym]
    y: [1x1 sym]
>> S.x
0
>> S.y
0
```

7.2 Eigenvectors by hand

8 Solutions

8.1 Subderivative of the hinge loss function

8.2 EM for the Student distribution

At first blush, it might not be apparent why EM can be used, since there is no missing data. The key idea is to introduce an "artificial" hidden or auxiliary variable in order to simplify the algorithm. In particular, we will exploit the fact that a Student distribution can be written as a **Gaussian scale mixture [Andrews74; West87]** as follows:

$$\mathcal{T}(\boldsymbol{x}_n|\boldsymbol{\mu},\boldsymbol{\Sigma},\nu) = \int \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu},\boldsymbol{\Sigma}/z_n)\mathrm{Ga}(z_n|\frac{\nu}{2},\frac{\nu}{2})dz_n \tag{138}$$

(See Exercise **??** for a proof of this in the 1d case.) This can be thought of as an "infinite" mixture of Gaussians, each one with a slightly different covariance matrix.

Treating the z_n as missing data, we can write the complete data log likelihood as

$$\ell_c(\boldsymbol{\theta}) = \sum_{n=1}^{N}\left[\log\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu},\boldsymbol{\Sigma}/z_n) + \log\mathrm{Ga}(z_n|\nu/2,\nu/2)\right] \tag{139}$$

$$= \sum_{n=1}^{N}\left[-\frac{D}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{z_n}{2}\delta_n + \frac{\nu}{2}\log\frac{\nu}{2} - \log\Gamma(\frac{\nu}{2})\right. \tag{140}$$

$$\left.+\frac{\nu}{2}(\log z_n - z_n) + (\frac{D}{2}-1)\log z_n\right] \tag{141}$$

where we have defined the Mahalanobis distance to be

$$\delta_n = (\boldsymbol{x}_n - \boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}) \tag{142}$$

We can partition this into two terms, one involving $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, and the other involving ν. We have, dropping irrelevant constants,

$$\ell_c(\boldsymbol{\theta}) = L_N(\boldsymbol{\mu},\boldsymbol{\Sigma}) + L_G(\nu) \tag{143}$$

$$L_N(\boldsymbol{\mu},\boldsymbol{\Sigma}) \triangleq -\frac{1}{2}N\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N} z_n\delta_n \tag{144}$$

$$L_G(\nu) \triangleq -N\log\Gamma(\nu/2) + \frac{1}{2}N\nu\log(\nu/2) + \frac{1}{2}\nu\sum_{n=1}^{N}(\log z_n - z_n) \tag{145}$$

Let us first derive the algorithm with ν assumed known, for simplicity. In this case, we can ignore the L_G term, so we only need to figure out how to compute $\mathbb{E}[z_n]$ wrt the old parameters.

From **??** we have

$$p(z_n|\boldsymbol{x}_n,\boldsymbol{\theta}) = \mathrm{Ga}(z_n|\frac{\nu+D}{2},\frac{\nu+\delta_n}{2}) \tag{146}$$

Now if $z_n \sim \mathrm{Ga}(a,b)$, then $\mathbb{E}[z_n] = a/b$. Hence the E step at iteration t is

$$\overline{z}_n^{(t)} \triangleq \mathbb{E}\left[z_n|\boldsymbol{x}_n,\boldsymbol{\theta}^{(t)}\right] = \frac{\nu^{(t)}+D}{\nu^{(t)}+\delta_n^{(t)}} \tag{147}$$

The M step is obtained by maximizing $\mathbb{E}\left[L_N(\boldsymbol{\mu}, \boldsymbol{\Sigma})\right]$ to yield

$$\hat{\boldsymbol{\mu}}^{(t+1)} = \frac{\sum_n \overline{z}_n^{(t)} \boldsymbol{x}_n}{\sum_n \overline{z}_n^{(t)}} \tag{148}$$

$$\hat{\boldsymbol{\Sigma}}^{(t+1)} = \frac{1}{N} \sum_n \overline{z}_n^{(t)} (\boldsymbol{x}_n - \hat{\boldsymbol{\mu}}^{(t+1)})(\boldsymbol{x}_n - \hat{\boldsymbol{\mu}}^{(t+1)})^\mathsf{T} \tag{149}$$

$$= \frac{1}{N} \left[\sum_n \overline{z}_n^{(t)} \boldsymbol{x}_n \boldsymbol{x}_n^\mathsf{T} - \left(\sum_{n=1}^N \overline{z}_n^{(t)} \right) \hat{\boldsymbol{\mu}}^{(t+1)} (\hat{\boldsymbol{\mu}}^{(t+1)})^\mathsf{T} \right] \tag{150}$$

These results are quite intuitive: the quantity $\overline{z}_n$ is the precision of measurement n, so if it is small, the corresponding data point is down-weighted when estimating the mean and covariance. This is how the Student achieves robustness to outliers.

To compute the MLE for the degrees of freedom, we first need to compute the expectation of $L_G(\nu)$, which involves z_n and $\log z_n$. Now if $z_n \sim \mathrm{Ga}(a, b)$, then one can show that

$$\overline{\ell}_n^{(t)} \triangleq \mathbb{E}\left[\log z_n | \boldsymbol{\theta}^{(t)}\right] = \psi(a) - \log b \tag{151}$$

where $\psi(x) = \frac{d}{dx} \log \Gamma(x)$ is the digamma function. Hence, from Equation (146), we have

$$\overline{\ell}_n^{(t)} = \Psi(\frac{\nu^{(t)} + D}{2}) - \log(\frac{\nu^{(t)} + \delta_n^{(t)}}{2}) \tag{152}$$

$$= \log(\overline{z}_n^{(t)}) + \Psi(\frac{\nu^{(t)} + D}{2}) - \log(\frac{\nu^{(t)} + D}{2}) \tag{153}$$

Substituting into Equation (145), we have

$$\mathbb{E}\left[L_G(\nu)\right] = -N \log \Gamma(\nu/2) + \frac{N\nu}{2} \log(\nu/2) + \frac{\nu}{2} \sum_n (\overline{\ell}_n^{(t)} - \overline{z}_n^{(t)}) \tag{154}$$

The gradient of this expression is equal to

$$\frac{d}{d\nu} \mathbb{E}\left[L_G(\nu)\right] = -\frac{N}{2} \Psi(\nu/2) + \frac{N}{2} \log(\nu/2) + \frac{N}{2} + \frac{1}{2} \sum_n (\overline{\ell}_n^{(t)} - \overline{z}_n^{(t)}) \tag{155}$$

This has a unique solution in the interval $(0, +\infty]$ which can be found using a 1d constrained optimizer.

Performing a gradient-based optimization in the M step, rather than a closed-form update, is an example of what is known as the **generalized EM** algorithm.

Part II
Linear models

9 Solutions

9.1 Derivation of Fisher's linear discriminant

We have

$$f = \boldsymbol{w}^T \mathbf{S}_B \boldsymbol{w} \tag{156}$$

$$f' = 2\mathbf{S}_B \boldsymbol{w} \tag{157}$$

$$g = \boldsymbol{w}^T \mathbf{S}_W \boldsymbol{w} \tag{158}$$

$$g' = 2\mathbf{S}_W \boldsymbol{w} \tag{159}$$

$$\frac{dJ(\boldsymbol{w})}{d\boldsymbol{w}} = \frac{(2\boldsymbol{s}_B \boldsymbol{w})(\boldsymbol{w}^T \mathbf{S}_W \boldsymbol{w}) - (\boldsymbol{w}^T \mathbf{S}_B \boldsymbol{w})(2\mathbf{S}_W \boldsymbol{w})}{(\boldsymbol{w}^T \mathbf{S}_W \boldsymbol{w})^T (\boldsymbol{w}^T \mathbf{S}_W \boldsymbol{w})} = 0 \tag{160}$$

Hence

$$(\boldsymbol{s}_B \boldsymbol{w}) \underbrace{(\boldsymbol{w}^T \mathbf{S}_W \boldsymbol{w})}_{a} = \underbrace{(\boldsymbol{w}^T \mathbf{S}_B \boldsymbol{w})}_{b} (\mathbf{S}_W \boldsymbol{w}) \tag{161}$$

$$a\mathbf{S}_B \boldsymbol{w} = b\mathbf{S}_W \boldsymbol{w} \tag{162}$$

$$\mathbf{S}_B \boldsymbol{w} = \frac{b}{a} \mathbf{S}_W \boldsymbol{w} \tag{163}$$

10 Solutions

10.1 Gradient and Hessian of log-likelihood for multinomial logistic regression

1. Let us drop the i subscript for simplicity. Let $S = \sum_k e^{\eta_k}$ be the denominator of the softmax.

$$\frac{\partial \mu_k}{\partial \eta_j} = \left[\frac{\partial}{\partial \eta_j} e^{\eta_k}\right] S^{-1} + e^{\eta_k} \cdot [\frac{\partial}{\partial \eta_j} S] \cdot -S^{-2} \tag{164}$$

$$= (\delta_{ij} e^{\eta_k}) S^{-1} - e^{\eta_k} e^{\eta_j} S^{-2} \tag{165}$$

$$= \frac{e^{\eta_k}}{S}\left(\delta_{ij} - \frac{e^{\eta_j}}{S}\right) \tag{166}$$

$$= \mu_k(\delta_{kj} - \mu_j) \tag{167}$$

2. We have

$$\nabla_{\boldsymbol{w}_j} \ell = \sum_i \sum_k \frac{\partial \ell}{\partial \mu_{ik}} \frac{\partial \mu_{ik}}{\partial \eta_{ij}} \frac{\partial \eta_{ij}}{\partial \boldsymbol{w}_j} = \sum_i \sum_k \frac{y_{ik}}{\mu_{ik}} \mu_{ik}(\delta_{jk} - \mu_{ij}) \boldsymbol{x}_i \tag{168}$$

$$= \sum_i \sum_k y_{ik}(\delta_{jk} - \mu_{ij}) \boldsymbol{x}_i = \sum_i y_{ij} \boldsymbol{x}_i - \sum_i (\sum_k y_{ik}) \mu_{ij} \boldsymbol{x}_i \tag{169}$$

$$= \sum_i (y_{ij} - \mu_{ij}) \boldsymbol{x}_i \tag{170}$$

3. We consider a single term $\boldsymbol{x}_i$ in the log likelihood; we can sum over i at the end. Using the Jacobian expression from above, we have

$$\nabla_{\boldsymbol{w}_{c'}} (\nabla_{\boldsymbol{w}_c} \ell)^T = \nabla_{\boldsymbol{w}_{c'}} ((y_{ic} - \mu_{ic}) \boldsymbol{x}_i^T) \tag{171}$$

$$= -(\nabla_{\boldsymbol{w}_{c'}} \mu_{ic}) \boldsymbol{x}_i^T \tag{172}$$

$$= -(\mu_{ic}(\delta_{c,c'} - \mu_{i,c'}) \boldsymbol{x}_i) \boldsymbol{x}_i^T \tag{173}$$

10.2 Regularizing separate terms in 2d logistic regression

10.3 Logistic regression vs LDA/QDA

11 Solutions

11.1 Multi-output linear regression

11.2 Centering and ridge regression

Suppose $\mathbf{X}$ is centered, so $\overline{\boldsymbol{x}} = 0$. Then

$$J(\boldsymbol{w}, w_0) = (\boldsymbol{y} - \mathbf{X}\boldsymbol{w} - w_0\mathbf{1})^T(\boldsymbol{y} - \mathbf{X}\boldsymbol{w} - w_0\mathbf{1}) + \lambda \boldsymbol{w}^T\boldsymbol{w} \tag{174}$$

$$= \boldsymbol{y}^T\boldsymbol{y} + \boldsymbol{w}^T(\mathbf{X}^T\mathbf{X})\boldsymbol{w} - 2\boldsymbol{y}^T(\mathbf{X}\boldsymbol{w}) + \lambda \boldsymbol{w}^T\boldsymbol{w} + \left(-2w_0\mathbf{1}^T\boldsymbol{y} + 2w_0\mathbf{1}^T\mathbf{X}\boldsymbol{w} + w_0\mathbf{1}^T\mathbf{1}w_0\right) \tag{175}$$

Consider the terms in brackets:

$$w_0\mathbf{1}^T\boldsymbol{y} = w_0 n\overline{y} \tag{176}$$

$$w_0\mathbf{1}^T\mathbf{X}\boldsymbol{w} = w_0 \sum_i \boldsymbol{x}_i^T\boldsymbol{w} = n\overline{\boldsymbol{x}}^T\boldsymbol{w} = 0 \tag{177}$$

$$w_0\mathbf{1}^T\mathbf{1}w_0 = nw_0^2 \tag{178}$$

Optimizing wrt w_0 we find

$$\frac{\partial}{\partial w_0} J(\boldsymbol{w}, w_0) = -2n\overline{y} + 2nw_0 = 0 \tag{179}$$

$$\hat{w}_0 = \overline{y} \tag{180}$$

Optimizing wrt $\boldsymbol{w}$ we find

$$\frac{\partial}{\partial \boldsymbol{w}} J(\boldsymbol{w}, \hat{w}_0) = [2\mathbf{X}^T\mathbf{X}\boldsymbol{w} - 2\mathbf{X}^T\boldsymbol{y}] + 2\lambda\boldsymbol{w} = 0 \tag{181}$$

$$\boldsymbol{w} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\boldsymbol{y} \tag{182}$$

11.3 Partial derivative of the RSS

11.4 Reducing elastic net to lasso

We have

$$J_1(\boldsymbol{w}) = \boldsymbol{y}^T\boldsymbol{y} + (\mathbf{X}\boldsymbol{w})^T(\mathbf{X}\boldsymbol{w}) - 2\boldsymbol{y}^T(\mathbf{X}\boldsymbol{w}) + \lambda_2\boldsymbol{w}^T\boldsymbol{w} + \lambda_1|\boldsymbol{w}|_1 \tag{183}$$

and

$$J_2(\boldsymbol{w}) = \boldsymbol{y}^T\boldsymbol{y} + c^2(\mathbf{X}\boldsymbol{w})^T(\mathbf{X}\boldsymbol{w}) - 2c^2\boldsymbol{y}^T(\mathbf{X}\boldsymbol{w}) + \lambda_2 c^2\boldsymbol{w}^T\boldsymbol{w} + c\lambda_1|\boldsymbol{w}|_1 = J_1(c\boldsymbol{w}) \tag{184}$$

11.5 Shrinkage in linear regression

11.6 EM for mixture of linear regression experts

In the E step, we compute the conditional responsibilities

$$r_{nk} = p(z_n = k|\boldsymbol{x}_n, \boldsymbol{y}_n) = \frac{p(\boldsymbol{y}_n|\boldsymbol{x}_n, z_n = k)p(z_n = k|\boldsymbol{x}_n)}{p(\boldsymbol{y}_n|\boldsymbol{x}_n)} \tag{185}$$

In the M step, we update the parameters of the gating funtion by maximizing the weighted likelihood

$$\ell(\boldsymbol{\theta}_g) = \sum_n \sum_k r_{nk} \log p(z_n = k|\boldsymbol{x}_n, \boldsymbol{\theta}_g) \tag{186}$$

and the parameters of the k'th expert by maximizing the weighted likelihood

$$\ell(\boldsymbol{\theta}_k) = \sum_n r_{nk} \log p(\boldsymbol{y}_n|\boldsymbol{x}_n, z_n = k, \boldsymbol{\theta}_k) \tag{187}$$

If the gating function and experts are linear models, these M steps correspond to convex subproblems that can be solved efficiently.

For example, consider a mixture of linear regression experts using logistic regression gating functions. In the M step, we need to maximize $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$ wrt $\boldsymbol{w}_k$, σ_k^2 and $\mathbf{V}$. For the regression parameters for model k, the objective has the form

$$Q(\boldsymbol{\theta}_k, \boldsymbol{\theta}^{old}) = \sum_{n=1}^{N} r_{nk} \left\{ -\frac{1}{\sigma_k^2}(y_n - \boldsymbol{w}_k^\mathsf{T}\boldsymbol{x}_n) \right\} \tag{188}$$

We recognize this as a weighted least squares problem, which makes intuitive sense: if r_{nk} is small, then data point n will be downweighted when estimating model k's parameters. From **??** we can immediately write down the MLE as

$$\boldsymbol{w}_k = (\mathbf{X}^\mathsf{T}\mathbf{R}_k\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{R}_k\boldsymbol{y} \tag{189}$$

where $\mathbf{R}_k = \text{diag}(r_{:,k})$. The MLE for the variance is given by

$$\sigma_k^2 = \frac{\sum_{n=1}^{N} r_{nk}(y_n - \boldsymbol{w}_k^\mathsf{T}\boldsymbol{x}_n)^2}{\sum_{n=1}^{N} r_{nk}} \tag{190}$$

We replace the estimate of the unconditional mixing weights $\boldsymbol{\pi}$ with the estimate of the gating parameters, $\mathbf{V}$. The objective has the form

$$\ell(\mathbf{V}) = \sum_n \sum_k r_{nk} \log \pi_{n,k} \tag{191}$$

We recognize this as equivalent to the log-likelihood for multinomial logistic regression in **??**, except we replace the "hard" 1-of-C encoding $\boldsymbol{y}_i$ with the "soft" 1-of-K encoding $\boldsymbol{r}_i$. Thus we can estimate $\mathbf{V}$ by fitting a logistic regression model to soft target labels.

12 Solutions

Part III
Deep neural networks

13 Solutions

13.1 Backpropagation for a 1 layer MLP

To compute $\boldsymbol{\delta}_1$, let $\mathcal{L} = \text{CrossEntropyWithLogits}(\boldsymbol{y}, \boldsymbol{a})$. Then

$$\boldsymbol{\delta}_1 = \frac{\partial \mathcal{L}}{\partial \boldsymbol{a}} = (\boldsymbol{p} - \boldsymbol{y})^\mathsf{T} \tag{192}$$

where $\boldsymbol{p} = \mathcal{S}(\boldsymbol{a})$.

To compute $\boldsymbol{\delta}_2$, we have

$$\boldsymbol{\delta}_2 = \frac{\partial \mathcal{L}}{\partial \boldsymbol{z}} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{a}} \frac{\partial \boldsymbol{a}}{\partial \boldsymbol{h}} \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{z}} = \boldsymbol{\delta}_1 \frac{\partial \boldsymbol{a}}{\partial \boldsymbol{h}} \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{z}} \tag{193}$$

$$= \boldsymbol{\delta}_1 \mathbf{U} \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{z}} \text{ since } \boldsymbol{a} = \mathbf{U}\boldsymbol{h} + \boldsymbol{b}_2 \tag{194}$$

$$= \boldsymbol{\delta}_1 \mathbf{U} \circ \text{ReLU}'(\boldsymbol{z}) \text{ since } \boldsymbol{h} = \text{ReLU}(\boldsymbol{z}) \tag{195}$$

$$= \boldsymbol{\delta}_1 \mathbf{U} \circ H(\boldsymbol{h}) \tag{196}$$

Now we compute the gradients wrt the parameters:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{U}} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{a}} \frac{\partial \boldsymbol{a}}{\partial \mathbf{U}} = \boldsymbol{\delta}_1 \boldsymbol{h}^\mathsf{T} \tag{197}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{b}_2} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{a}} \frac{\partial \boldsymbol{a}}{\partial \boldsymbol{b}_2} = \boldsymbol{\delta}_1 \tag{198}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{z}} \frac{\partial \boldsymbol{z}}{\partial \mathbf{W}} = \boldsymbol{\delta}_2 \boldsymbol{x}^\mathsf{T} \tag{199}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{b}_1} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{z}} \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{b}_1} = \boldsymbol{\delta}_2 \tag{200}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{x}} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{z}} \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}} = \boldsymbol{\delta}_2 \mathbf{W} \tag{201}$$

14 Solutions

15 Solutions

Part IV
Nonparametric models

16 Solutions

17 Solutions

17.1 Fitting an SVM classifier by hand

18 Solutions

Part V
Beyond supervised learning

19 Solutions

19.1 Information gain equations

To see this, let us define $p_n \triangleq p(\boldsymbol{\theta}|\mathcal{D})$ as the current belief state, and $p_{n+1} \triangleq p(\boldsymbol{\theta}|\mathcal{D}, \boldsymbol{x}, y)$ as the belief state after observing $(\boldsymbol{x}, y)$. Then

$$\mathbb{KL}\left(p_{n+1} \| p_n\right) = \int p_{n+1} \log \frac{p_{n+1}}{p_n} d\boldsymbol{\theta} \tag{202}$$

$$= \int p_{n+1} \log p_{n+1} d\boldsymbol{\theta} - \int p_{n+1} \log p_n d\boldsymbol{\theta} \tag{203}$$

$$= -\mathbb{H}\, p_{n+1} - \int p_{n+1} \log p_n d\boldsymbol{\theta} \tag{204}$$

Next we need to take expectations wrt $p(y|\boldsymbol{x}, \mathcal{D})$. We will use the following identity:

$$\sum_y p(y|\boldsymbol{x}, \mathcal{D}) \int p_{n+1} \log p_n d\boldsymbol{\theta} \tag{205}$$

$$= \sum_y p(y|\boldsymbol{x}, \mathcal{D}) \int p(\boldsymbol{\theta}|\mathcal{D}, \boldsymbol{x}, y) \log p(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta} \tag{206}$$

$$= \int \sum_y p(y, \boldsymbol{\theta}|\boldsymbol{x}, \mathcal{D}) \log p(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta} \tag{207}$$

$$= \int p(\boldsymbol{\theta}|\mathcal{D}) \log p(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta} \tag{208}$$

Using this, we have

$$U'(\boldsymbol{x}) = \mathbb{E}_{p(y|\boldsymbol{x},\mathcal{D})}\left[-\mathbb{H}\, p_{n+1}\right] - \int p_n \log p_n d\boldsymbol{\theta} \tag{209}$$

$$= \mathbb{E}_{p(y|\boldsymbol{x},\mathcal{D})}\left[-\mathbb{H}\, p(\boldsymbol{\theta}|\boldsymbol{x}, y, \mathcal{D}) + \mathbb{H}\, p(\boldsymbol{\theta}|\mathcal{D})\right] = U(\boldsymbol{x}) \tag{210}$$

20 Solutions

20.1 EM for FA

In the E step, we can compute the posterior for $\boldsymbol{z}_n$ as follows:

$$p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{z}_n|\boldsymbol{m}_n, \boldsymbol{\Sigma}_n) \tag{211}$$

$$\boldsymbol{\Sigma}_n \triangleq (\mathbf{I} + \mathbf{W}^\mathsf{T}\boldsymbol{\Psi}^{-1}\mathbf{W})^{-1} \tag{212}$$

$$\boldsymbol{m}_n \triangleq \boldsymbol{\Sigma}_n(\mathbf{W}^\mathsf{T}\boldsymbol{\Psi}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu})) \tag{213}$$

We now discuss the M step. We initially assume $\boldsymbol{\mu} = \mathbf{0}$. Using the trace trick we have

$$\sum_i \mathbb{E}\left[(\tilde{\mathbf{x}}_i - \mathbf{W}\boldsymbol{z}_i)^T\boldsymbol{\Psi}^{-1}(\tilde{\mathbf{x}}_i - \mathbf{W}\boldsymbol{z}_i)\right] = \sum_i \left[\tilde{\mathbf{x}}_i^T\boldsymbol{\Psi}^{-1}\tilde{\mathbf{x}}_i + \mathbb{E}\left[\boldsymbol{z}_i^T\mathbf{W}^T\boldsymbol{\Psi}^{-1}\mathbf{W}\boldsymbol{z}_i\right] - 2\tilde{\mathbf{x}}_i^T\boldsymbol{\Psi}^{-1}\mathbf{W}\mathbb{E}\left[\boldsymbol{z}_i\right]\right] \tag{214}$$

$$= \sum_i \Big[\mathrm{tr}(\boldsymbol{\Psi}^{-1}\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^T) + \mathrm{tr}(\boldsymbol{\Psi}^{-1}\mathbf{W}\mathbb{E}\left[\boldsymbol{z}_i\boldsymbol{z}_i^T\right]\mathbf{W}^T) \tag{215}$$

$$-\mathrm{tr}(2\boldsymbol{\Psi}^{-1}\mathbf{W}\mathbb{E}\left[\boldsymbol{z}_i\right]\tilde{\mathbf{x}}_i^T)\Big] \tag{216}$$

$$\triangleq \mathrm{tr}(\boldsymbol{\Psi}^{-1}\mathbf{G}(\mathbf{W})) \tag{217}$$

Hence the expected complete data log likelihood is given by

$$Q = \frac{N}{2}\log|\boldsymbol{\Psi}^{-1}| - \frac{1}{2}\mathrm{tr}(\boldsymbol{\Psi}^{-1}\mathbf{G}(\mathbf{W})) \tag{218}$$

Using the chain rule and the facts that $\frac{\partial}{\partial\mathbf{W}}\mathrm{tr}(\mathbf{W}^T\mathbf{A}) = \mathbf{A}$ $\frac{\partial}{\partial\mathbf{W}}\mathbf{W}^T\mathbf{A}\mathbf{W} = (\mathbf{A} + \mathbf{A}^T)\mathbf{W}$ we have

$$\nabla_{\mathbf{W}}Q(\mathbf{W}) = -\frac{1}{2}\boldsymbol{\Psi}^{-1}\nabla_{\mathbf{W}}G(\mathbf{W}) = 0 \tag{219}$$

$$\nabla_{\mathbf{W}}G(\mathbf{W}) = 2\mathbf{W}\sum_i \mathbb{E}\left[\boldsymbol{z}_i\boldsymbol{z}_i^T\right] - 2(\sum_i \mathbb{E}\left[\boldsymbol{z}_i\right]\tilde{\mathbf{x}}_i^T)^T \tag{220}$$

$$\mathbf{W}_{mle} = \left[\sum_i \tilde{\mathbf{x}}_i\mathbb{E}\left[\boldsymbol{z}_i\right]^T\right]\left[\sum_i \mathbb{E}\left[\boldsymbol{z}_i\boldsymbol{z}_i^T\right]\right]^{-1} \tag{221}$$

Using the facts that $\nabla_{\mathbf{X}}\log|\mathbf{X}| = \mathbf{X}^{-T}$ and $\nabla_{\mathbf{X}}\mathrm{tr}(\mathbf{X}\mathbf{A}) = \mathbf{A}^T$ we have

$$\nabla_{\boldsymbol{\Psi}^{-1}}Q = \frac{N}{2}\boldsymbol{\Psi} - \frac{1}{2}\mathbf{G}(\mathbf{W}_{mle}) = 0 \tag{222}$$

$$\boldsymbol{\Psi} = \frac{1}{N}\mathrm{diag}(\mathbf{G}(\mathbf{W}_{mle})) \tag{223}$$

We can simplify this as follows, by plugging in the MLE (this simplification no longer holds if we use MAP estimation). First note that

$$\sum_i \mathbb{E}\left[\boldsymbol{z}_i\boldsymbol{z}_i^T\right]\mathbf{W}_{mle}^T = \sum_i \mathbb{E}\left[\boldsymbol{z}_i\right]\tilde{\mathbf{x}}_i^T \tag{224}$$

so

$$\boldsymbol{\Psi} = \frac{1}{N}\sum_i \left(\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^T + \mathbf{W}_{mle}\mathbb{E}\left[\boldsymbol{z}_i\right]\tilde{\mathbf{x}}_i^T - 2\mathbf{W}_{mle}\mathbb{E}\left[\boldsymbol{z}_i\right]\tilde{\mathbf{x}}_i^T\right) \tag{225}$$

$$= \frac{1}{N}\left(\sum_i \tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^T - \mathbf{W}_{mle}\sum_i \mathbb{E}\left[\boldsymbol{z}_i\right]\tilde{\mathbf{x}}_i^T\right) \tag{226}$$

To estimate $\boldsymbol{\mu}$ and $\mathbf{W}$ at the same time, we can define $\tilde{\mathbf{W}} = (\mathbf{W}, \boldsymbol{\mu})$ and $\tilde{\boldsymbol{z}} = (\boldsymbol{z}, 1)$, Also, define

$$\boldsymbol{b}_n \triangleq \mathbb{E}\left[\tilde{\mathbf{z}}|\boldsymbol{x}_n\right] = [\boldsymbol{m}_n; 1] \tag{227}$$

$$\boldsymbol{\Omega}_n \triangleq \mathbb{E}\left[\tilde{\mathbf{z}}\tilde{\mathbf{z}}^{\mathsf{T}}|\boldsymbol{x}_n\right] = \begin{pmatrix} \mathbb{E}\left[\boldsymbol{z}\boldsymbol{z}^{\mathsf{T}}|\boldsymbol{x}_n\right] & \mathbb{E}\left[\boldsymbol{z}|\boldsymbol{x}_n\right] \\ \mathbb{E}\left[\boldsymbol{z}|\boldsymbol{x}_n\right]^{\mathsf{T}} & 1 \end{pmatrix} \tag{228}$$

Then the M step is as follows:

$$\hat{\tilde{\mathbf{W}}} = \left[\sum_n \boldsymbol{x}_n \boldsymbol{b}_n^{\mathsf{T}}\right]\left[\sum_n \boldsymbol{\Omega}_n\right]^{-1} \tag{229}$$

$$\hat{\boldsymbol{\Psi}} = \frac{1}{N}\mathrm{diag}\left\{\sum_n \left(\boldsymbol{x}_n - \hat{\tilde{\mathbf{W}}}\boldsymbol{b}_n\right)\boldsymbol{x}_n^{\mathsf{T}}\right\} \tag{230}$$

It is interesting to apply the above equations to the PPCA case in the limit where $\sigma^2 \to 0$. This provides an alternative way to fit PCA models, as shown by [**Roweis97**]. Let $\tilde{\mathbf{Z}}$ be a $L \times N$ matrix storing the posterior means (low-dimensional representations) along its columns. Similarly, let $\tilde{\mathbf{X}} = \mathbf{X}^{\mathsf{T}}$ be an $D \times N$ matrix storing the original data along its columns. From **??**, when $\sigma^2 = 0$, we have

$$\tilde{\mathbf{Z}} = (\mathbf{W}^{\mathsf{T}}\mathbf{W})^{-1}\mathbf{W}^{\mathsf{T}}\tilde{\mathbf{X}} \tag{231}$$

This constitutes the E step. Notice that this is just an orthogonal projection of the data.

From Equation (229), the M step is given by

$$\hat{\mathbf{W}} = \left[\sum_n \boldsymbol{x}_n \mathbb{E}\left[\boldsymbol{z}_n\right]^{\mathsf{T}}\right]\left[\sum_n \mathbb{E}\left[\boldsymbol{z}_n\right]\mathbb{E}\left[\boldsymbol{z}_n\right]^{\mathsf{T}}\right]^{-1} \tag{232}$$

where we exploited the fact that $\boldsymbol{\Sigma} = \mathrm{Cov}\left[\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta}\right] = 0\mathbf{I}$ when $\sigma^2 = 0$. It is worth comparing this expression to the MLE for multi-output linear regression (**??**), which has the form $\mathbf{W} = (\sum_n \boldsymbol{y}_n \boldsymbol{x}_n^{\mathsf{T}})(\sum_n \boldsymbol{x}_n \boldsymbol{x}_n^{\mathsf{T}})^{-1}$. Thus we see that the M step is like linear regression where we replace the observed inputs by the expected values of the latent variables.

In summary, here is the entire algorithm:

- E step: $\tilde{\mathbf{Z}} = (\mathbf{W}^{\mathsf{T}}\mathbf{W})^{-1}\mathbf{W}^{\mathsf{T}}\tilde{\mathbf{X}}$
- M step: $\mathbf{W} = \tilde{\mathbf{X}}\tilde{\mathbf{Z}}^{\mathsf{T}}(\tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^{\mathsf{T}})^{-1}$

[**Tipping99b**] showed that the only stable fixed point of the EM algorithm is the globally optimal solution. That is, the EM algorithm converges to a solution where $\mathbf{W}$ spans the same linear subspace as that defined by the first L eigenvectors. However, if we want $\mathbf{W}$ to be orthogonal, and to contain the eigenvectors in descending order of eigenvalue, we have to orthogonalize the resulting matrix (which can be done quite cheaply). Alternatively, we can modify EM to give the principal basis directly [**Ahn03**].

20.2 EM for mixFA

20.3 Deriving the second principal component

1. Dropping terms that no not involve $\boldsymbol{z}_2$ we have

$$J = \frac{1}{n}\sum_{i=1}^{n}\left[-2z_{i2}\boldsymbol{v}_2^T(\boldsymbol{x}_i - z_{i1}\boldsymbol{v}_1) + z_{i2}^2\boldsymbol{v}_2^T\boldsymbol{v}_2\right] = \frac{1}{n}\sum_{i=1}^{n}\left[-2z_{i2}\boldsymbol{v}_2^T\boldsymbol{x}_i + z_{i2}^2\right] \tag{233}$$

since $\boldsymbol{v}_2^T \boldsymbol{v}_2 = 1$ and $\boldsymbol{v}_1^T \boldsymbol{v}_2 = 0$. Hence

$$\frac{\partial J}{\partial z_{i2}} = -2\boldsymbol{v}_2^T \boldsymbol{x}_i + 2z_{i2} = 0 \tag{234}$$

so

$$z_{i2} = \boldsymbol{v}_2^T \boldsymbol{x}_i \tag{235}$$

2. We have

$$\frac{\partial \tilde{J}}{\partial \boldsymbol{v}_2} = -2\mathbf{C}\boldsymbol{v}_2 + 2\lambda_2 \boldsymbol{v}_2 + \lambda_{12}\boldsymbol{v}_1 = 0 \tag{236}$$

Premultiplying by $\boldsymbol{v}_1^T$ yields

$$0 = -2\boldsymbol{v}_1^T \mathbf{C}\boldsymbol{v}_2 + 2\lambda_2 \boldsymbol{v}_1^T \boldsymbol{v}_2 + \lambda_{12}\boldsymbol{v}_1^T \boldsymbol{v}_1 \tag{237}$$

Now $\boldsymbol{v}_1^T \mathbf{C}\boldsymbol{v}_2 = \boldsymbol{v}_1^T(\lambda_1 \boldsymbol{v}_2) = 0$, and $\boldsymbol{v}_1^T \boldsymbol{v}_2 = 0$, and $\boldsymbol{v}_1^T \boldsymbol{v}_1 = 1$, so $\lambda_{12} = 0$. Hence

$$0 = -2\mathbf{C}\boldsymbol{v}_2 + 2\lambda_2 \boldsymbol{v}_2 \tag{238}$$

$$\mathbf{C}\boldsymbol{v}_2 = \lambda_2 \boldsymbol{v}_2 \tag{239}$$

So $\boldsymbol{v}_2$ is an eigenvector of $\mathbf{C}$. Since we want to maximize the variance, we want to pick the eigenvector with the largest eigenvalue, but the first one is already taken. Hence $\boldsymbol{v}_2$ is the evector with the second largest evalue.

20.4 Deriving the residual error for PCA

20.5 PCA via successive deflation

1. We have

$$\tilde{\mathbf{C}} = \frac{1}{n}\left[(\mathbf{I} - \boldsymbol{v}_1 \boldsymbol{v}_1^T)\mathbf{X}^T \mathbf{X}(\mathbf{I} - \boldsymbol{v}_1 \boldsymbol{v}_1^T)\right] \tag{240}$$

$$= \frac{1}{n}\left[(\mathbf{X}^T\mathbf{X} - \boldsymbol{v}_1 \boldsymbol{v}_1^T \mathbf{X}^T\mathbf{X})(\mathbf{I} - \boldsymbol{v}_1 \boldsymbol{v}_1^T)\right] \tag{241}$$

$$= \frac{1}{n}\left[\mathbf{X}^T\mathbf{X} - \boldsymbol{v}_1(\boldsymbol{v}_1^T \mathbf{X}^T\mathbf{X}) - (\mathbf{X}^T\mathbf{X}\boldsymbol{v}_1)\boldsymbol{v}_1^T + \boldsymbol{v}_1(\boldsymbol{v}_1^T \mathbf{X}^T\mathbf{X}\boldsymbol{v}_1)\boldsymbol{v}_1^T\right] \tag{242}$$

$$= \frac{1}{n}\left[\mathbf{X}^T\mathbf{X} - \boldsymbol{v}_1(n\lambda_1 \boldsymbol{v}_1^T) - (n\lambda_1 \boldsymbol{v}_1)\boldsymbol{v}_1^T + \boldsymbol{v}_1(\boldsymbol{v}_1^T n\lambda_1 \boldsymbol{v}_1)\boldsymbol{v}_1^T\right] \tag{243}$$

$$= \frac{1}{n}\left[\mathbf{X}^T\mathbf{X} - n\lambda_1 \boldsymbol{v}_1 \boldsymbol{v}_1^T - n\lambda_1 \boldsymbol{v}_1 \boldsymbol{v}_1^T + n\lambda_1 \boldsymbol{v}_1 \boldsymbol{v}_1^T\right] \tag{244}$$

$$= \frac{1}{n}\mathbf{X}^T\mathbf{X} - \lambda_1 \boldsymbol{v}_1 \boldsymbol{v}_1^T \tag{245}$$

2. Since $\tilde{\mathbf{X}}$ lives in the $d-1$ subspace orthogonal to $\boldsymbol{v}_1$, the vector $\boldsymbol{u}$ must be orthogonal to $\boldsymbol{v}_1$. Hence $\boldsymbol{u}^T \boldsymbol{v}_1 = 0$ and $\boldsymbol{u}^T \boldsymbol{u} = 1$, so $\boldsymbol{u} = \boldsymbol{v}_2$.

3. We have

```
function [V, lambda] = simplePCA(C, K, f)
d = length(C);
V = zeros(d,K);
for j=1:K
  [lambda(j), V(:,j) = f(C);
  C = C - lambda(j)*V(:,j)*V(:,j)'; % deflation
end
```

20.6 PPCA variance terms

Define $\mathbf{A} = (\mathbf{\Lambda}_K - \sigma^2\mathbf{I})^{\frac{1}{2}}$.

1. We have

$$\boldsymbol{v}^T\mathbf{C}\boldsymbol{v} = \boldsymbol{v}^T(\mathbf{UARR}^T\mathbf{A}^T\mathbf{U}^T + \sigma^2\mathbf{I})\boldsymbol{v} = \boldsymbol{v}^T\sigma^2\mathbf{I}\boldsymbol{v} = \sigma^2 \tag{246}$$

2. We have

$$\begin{aligned}\boldsymbol{v}^T\mathbf{C}\boldsymbol{v} &= \boldsymbol{u}_i^T(\mathbf{UARR}^T\mathbf{A}^T\mathbf{U}^T + \sigma^2\mathbf{I})\boldsymbol{u}_i &(247)\\ &= \boldsymbol{u}_i^T\mathbf{U}(\mathbf{\Lambda} - \sigma^2\mathbf{I})\mathbf{U}^T\boldsymbol{u}_i + \sigma^2 &(248)\\ &= \boldsymbol{e}_i^T(\mathbf{\Lambda} - \sigma^2\mathbf{I})\boldsymbol{e}_i + \sigma^2 = \lambda_i - \sigma^2 + \sigma^2 &(249)\end{aligned}$$

20.7 Posterior inference in PPCA

20.8 Imputation in a FA model

20.9 Efficiently evaluating the PPCA density

Since $\mathbf{C}$ is not full rank, we can use matrix inversion lemma to invert it efficiently:

$$\mathbf{C}^{-1} = \frac{1}{\sigma^2}\left[\mathbf{I} - \mathbf{W}\left(\mathbf{W}^T\mathbf{W} + \sigma^2\mathbf{I}\right)\mathbf{W}^T\right] \tag{250}$$

Plugging in the MLE we find

$$\begin{aligned}\mathbf{W} &= \mathbf{U}_K(\mathbf{\Lambda}_K - \sigma^2\mathbf{I})^{\frac{1}{2}} &(251)\\ \mathbf{C}^{-1} &= \frac{1}{\sigma^2}\left[\mathbf{I} - \mathbf{U}_K(\mathbf{\Lambda}_K - \sigma^2\mathbf{I})\mathbf{\Lambda}_K^{-1}\mathbf{U}_K^T\right] &(252)\\ &= \frac{1}{\sigma^2}\left[\mathbf{I} - \mathbf{U}_K\mathbf{J}\mathbf{U}_K^T\right] &(253)\\ \mathbf{J} &= \mathrm{diag}(1 - \sigma^2/\lambda_j) &(254)\end{aligned}$$

Similarly it can be shown that

$$\log|\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}| = (d-K)\log\sigma^2 + \sum_{i=1}^{K}\log\lambda_i \tag{255}$$

21 Solutions

22 Solutions

23 Solutions