COURSE 02457

Non-Linear Signal Processing: Quiz Exercise 7

C.M. Bishop: Pattern Recognition and Machine Learning, Sections 2.3 (especially 2.3.4), 2.3.9, 9, 9.2-9.3.1

Questions

1. The following is true of the maximum likelihood (ML) estimates of the mean and covariance matrix of a Gaussian distribution

   (a) The ML estimate of the mean is biased, and the ML estimate of the covariance matrix is also biased.

   (b) The ML estimate of the mean is biased, but the ML estimate of the covariance matrix is not biased.

   (c) The ML estimate of the mean is not biased, but the ML estimate of the covariance matrix is biased.

   (d) The ML estimate of the mean is not biased, and the ML estimate of the covariance matrix is not biased either.

2. What can be said of the Gaussian distribution compared to Student's t-distribution?

   (a) The t-distribution and the Gaussian distribution have equally many parameters.

   (b) The t-distribution has fewer parameters than the Gaussian distribution.

   (c) The t-distribution has thicker tails than the Gaussian distribution.

   (d) The t-distribution has thinner tails than the Gaussian distribution.

3. How many parameters (not independent) appear in a Gaussian mixture model with K components when the dimension of the space is one?

   (a) 3

   (b) K

   (c) $3+K$

   (d) $3K$

4. In Gaussian mixture models, the term "responsibility" refers to

   (a) The probability that a certain cluster generated a data point, i.e. the probability that the data point belongs to that cluster.

   (b) The probability that an arbitrary data point belongs to a cluster

   (c) The log-likelihood of the fit of the model to data.

   (d) The negative log-likelihood of the fit of the model to data.

5. Consider equation (2.44) in the textbook, which gives the expression in the exponent of the Gaussian distribution. Let $\mathbf{X} = (X_1, X_2)$ be a random variable that follows the Gaussian distribution with mean zero and inverse covariance matrix

$$\Sigma^{-1} = \begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{pmatrix}.$$

Give an expression in x_1 and x_2 such that the joint probability density $f(X_1 = x_1, X_2 = x_2)$ is constant when that expression is constant.

(a) $\lambda_{11}x_1^2 + \lambda_{22}x_2^2$
(b) $\lambda_{11}x_1^2 + \lambda_{22}x_2^2 + 2\lambda_{12}x_1x_2$
(c) $\lambda_{11}x_1^2 + \lambda_{22}x_2^2 + \lambda_{12}x_1x_2 + \lambda_{21}x_1x_2$
(d) $\lambda_{11}x_1^2 + \lambda_{22}x_2^2 - \lambda_{12}x_1x_2 - \lambda_{21}x_1x_2$

6. Let two Gaussian random variables X and Y have joint density $(X, Y) \sim \mathcal{N}(\mu, \Sigma)$ with

$$\mu = (1, -2)$$
$$\Sigma = \begin{pmatrix} 2/3 & 1/3 \\ 1/3 & 2/3 \end{pmatrix}.$$

What are μ and σ^2 in the conditional density $p(X = x|Y = y) \sim \mathcal{N}(\mu, \sigma)$?

(a) $\mu = 1 - \frac{1}{2}(y + 2)$, $\sigma^2 = 3/2$
(b) $\mu = 1 - \frac{1}{2}(y + 2)$, $\sigma^2 = 1/2$
(c) $\mu = 1 + \frac{1}{2}(y + 2)$, $\sigma^2 = 3/2$
(d) $\mu = 1 + \frac{1}{2}(y + 2)$, $\sigma^2 = 1/2$

7. What is the result of evaluating $\int \exp\left(-\frac{1}{2}(\mathbf{x}_b - \mathbf{\Lambda}_{bb}^{-1}\mathbf{m})^T\mathbf{\Lambda}_{bb}(\mathbf{x}_b - \mathbf{\Lambda}_{bb}^{-1}\mathbf{m})\right) d\mathbf{x}_b$? Let D denote the dimensions of $\mathbf{\Lambda}_{bb}$.

(a) The integral can only be approximated numerically
(b) 1
(c) $\left(\frac{1}{(2\,\pi)^{D/2}}\,\frac{1}{|\mathbf{\Lambda}_{bb}^{-1}|^{1/2}}\right)^{-1}$
(d) $\frac{1}{(2\,\pi)^{D/2}}\,\frac{1}{|\mathbf{\Lambda}_{bb}^{-1}|^{1/2}}$

8. Let $Z = (X, Y)$ be a random vector. If we know $p(X = x)$ and $p(Y = y|X = x)$, what is an expression for the log-density of Z?

(a) $\log(p(Z = z)) = \exp(p(X = x)) + \exp(p(Y = y|X = x))$
(b) $\log(p(Z = z)) = \log(p(X = x) + p(Y = y|X = x))$
(c) $\log(p(Z = z)) = \log(p(X = x)) + \log(p(Y = y|X = x))$
(d) $\log(p(Z = z)) = \log(p(X = x)) \cdot \log(p(Y = y|X = x))$

9. Let two Gaussian random variables X and Y have joint density $(X, Y) \sim \mathcal{N}(\mu, \Sigma)$ with

$$\mu = (5,\ 3)$$
$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 10 \end{pmatrix}.$$

What are the expected maximum-likelihood estimates $\hat{\mu}$ and $\hat{\Sigma}$ of the mean and covariance of the joint distribution based on 15 samples from this distribution?

(a) $\hat{\mu} = \frac{14}{15}\mu$, $\hat{\Sigma} = \frac{14}{15}\Sigma$

(b) $\hat{\mu} = \mu$, $\hat{\Sigma} = \frac{1}{15}\Sigma$

(c) $\hat{\mu} = \mu$, $\hat{\Sigma} = \frac{14}{15}\Sigma$

(d) $\hat{\mu} = \mu$, $\hat{\Sigma} = \frac{1}{14}\Sigma$

10. Let two Gaussian random variables X and Y have joint density $(X, Y) \sim \mathcal{N}(\mu, \Sigma)$ with

$$\mu = (5,\ 3)$$
$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 10 \end{pmatrix}.$$

Which of the following statements is true?

(a) X and Y are independent. From the expected maximum-likelihood estimates, X and Y would also be judged independent.

(b) X and Y are not independent. From the expected maximum-likelihood estimates, X and Y would be judged independent.

(c) X and Y are independent. However, from the expected maximum-likelihood estimates, X and Y would not be judged independent.

(d) X and Y are not independent. From the expected maximum-likelihood estimates, X and Y would not be judged independent either.

11. Consider the log-likelihood of the Gaussian mixture distribution

$$\ln(p(\mathbf{X}|\pi, \mu, \Sigma)) = \sum_{n=1}^{N} \ln(\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k)).$$

Why can an optimal solution for this not be found analytically when $K > 1$?

(a) An optimal solution cannot be found analytically for $K = 1$ either.

(b) Because the logarithm of a sum cannot be expanded meaning that the logarithm does not work directly on an exponential.

(c) Since there are two sums, a gridsearch has to be performed to find the optimal parameters.

(d) Because several probabilities have to be evaluated for each observation.

12. How many latent variables of the form $\mathbf{z} = (0, 0, \ldots, 1, 0, \ldots)$ with K entries are there in a Gaussian mixture model with K Gaussian distributions and N observations?

(a) N

(b) K

(c) $N \cdot K$

(d) $N + K$

13. Which of the following statements is true?

(a) K-means is a discriminative function and the Gaussian mixture model is a discriminative model

(b) K-means is a discriminative model and the Gaussian mixture model is a generative model

(c) K-means is a discriminative function and the Gaussian mixture model is a generative model

(d) Both K-means and the Gaussian mixture model are generative models

14. In the EM algorithm, E stands for expectation and M stands for maximization. When the EM algorithm is used to optimize Gaussian mixture models, what is updated in each step?

(a) For each observation, the probability of cluster membership is evaluated in the E step for each cluster. The effective numbers of points in each cluster are also updated in the E step. Means and covariances are updated in the M step.

(b) Means and covariances are updated in the E step. For each observation, the probability of cluster membership is evaluated in the M step for each cluster. The effective numbers of points in each cluster are also updated in the M step.

(c) Means, covariances, and effective numbers of points in each cluster are updated in the E step. For each observation, the probability of cluster membership is evaluated in the M step for each cluster.

(d) For each observation, the probability of cluster membership is evaluated in the E step for each cluster. Means, covariances, and effective numbers of points in each cluster are updated in the M step.

15. Assume we have three observations of a univariate random variable $\mathbf{x} = (-2, 2, 7)$ that wewant to fit with a Gaussian mixture model with two clusters. Assume also that we initialize the means of the clusters as $\mu_1 = 1$ and $\mu_2 = 3$, the standard deviations of the clusters as $\sigma_1 = \sigma_2 = 2$ and the π parameters as $\pi_1 = \pi_2 = 1/2$. After initialization and two iterations, what is the estimated probability that the observation $x_2 = 2$ belongs to cluster two ($\gamma(z_{22})$)? What are the estimated cluster means and standard deviations?

(a) $\gamma(z_{22})$=0.462, $\boldsymbol{\mu} = (-0.386, 4.96)$, $\boldsymbol{\sigma} = (2.15, 2.85)$

(b) $\gamma(z_{22})$=0.462, $\boldsymbol{\mu} = (-0.386, 4.96)$, $\boldsymbol{\sigma} = (2.15^2, 2.85^2)$

(c) $\gamma(z_{22})$=0.538, $\boldsymbol{\mu} = (-0.386, 4.96)$, $\boldsymbol{\sigma} = (2.15, 2.85)$

(d) $\gamma(z_{22})$=0.538, $\boldsymbol{\mu} = (-0.386, 4.96)$, $\boldsymbol{\sigma} = (2.15^2, 2.85^2)$

16. Challenge. Warning requires reading of Section 2.3.5 in Bishop! Consider the Robbins-Monro procedure for sequential estimation of the root of a function $f(\theta) = E(z|\theta)$, where values of z are observed (the root of a functions is the value of the independent variable for which the function is zero). The procedure is described in section 2.3.5 in Bishop and briefly restated here. The iteration step is $\theta^{(N)} = \theta^{(N-1)} + a_{N-1} \cdot z(\theta^{(N-1)})$ (equation (2.129) in Bishop). The coefficients $\{a_N\}$ are positive numbers that satisfy certain conditions given in section 2.3.5. The estimate of the root at step N is $\theta^{(N)}$ and $z(\theta^{(N)})$ is an observation of the dependent (response) variable when θ takes the value θ^N.

    If we want to use this procedure to find the ML estimate of a parameter θ, where the observed variables are $\{x_n | n = 1, \ldots, N\}$, what should we use as the function $z(\theta)$?

    (a) $z(\theta^{N-1}) = \frac{\partial \ln(p(x_N|\theta^{N-1}))}{\partial \theta^{N-1}}$

    (b) $z(\theta^{N-1}) = \frac{\partial -\ln(p(x_N|\theta^{N-1}))}{\partial \theta^{N-1}}$

    (c) $z(\theta^{N-1}) = -\ln(p(x_N|\theta^{N-1}))$

    (d) $z(\theta^{N-1}) = \ln(p(x_N|\theta^{N-1}))$

Hint - reading for each question

1. Section 2.3.4
2. Section 2.3.7
3. Section 2.3.9
4. Section 9.2, page 432
5. Section 2.3
6. Section 2.3.1
7. Section 2.3
8. Section 2.3.3
9. Section 2.3.4
10. Section 2.3.4
11. Section 2.3.9
12. Section 9.2
13. Sections 9.1, 9.2, and 1.5.4
14. Section 9.2.2

15. Section 9.2.2

16. Section 2.3.5

Correct answers

1. (c) For a random vector $\mathbf{X}$ following a normal distribution $\mathcal{N}(\mu, \Sigma)$, the expected value of the maximum likelihood estimate of the mean is μ, and so is unbiased. The expected value of the maximum likelihood for the covariance matrix is $\frac{N-1}{N}\Sigma$, and thus biased. The variable N denotes the number of observations on which the estimate is based.

2. (c) The Gaussian distribution has two parameters, the mean and the variance. Student's t-distribution has one additional parameter, the number of degrees of freedom. This excludes both answers (a) and (b). The t-distribution has thicker tails than the normal distribution, so answer (c) is correct.

3. (d) When the dimension of the space is one, meaning that the observations to be modeled are univariate, the mean, standard deviation, and the parameter π must be estimated for each cluster. With three parameters for each of K clusters, the total number of parameters to be estimated is $3K$.

4. (a) By "responsibility", we mean the probability $\gamma(z_k) = p(z_k = 1|\mathbf{x})$. This is the probability that the generating cluster for the observation was cluser k given the observation $\mathbf{x}$. In other words, the probability $\gamma(z_k)$, referred to as "responsibility", is the probability that a certain observation was generated by a certain cluster.

5. (b) and (c) From equation (2.44), we have the expression in the exponent in the Gaussian density:

$$\begin{aligned}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) =& (x_1-\mu_1 \quad x_2-\mu_2)\begin{pmatrix}\lambda_{11} & \lambda_{12}\\ \lambda_{21} & \lambda_{22}\end{pmatrix}\begin{pmatrix}x_1-\mu_1\\ x_2-\mu_2\end{pmatrix}\\ =&(x_1-\mu_1)[\lambda_{11}(x_1-\mu_1)+\lambda_{12}(x_2-\mu_2)]+\\ &(x_2-\mu_2)[\lambda_{21}(x_1-\mu_1)+\lambda_{22}(x_2-\mu_2)].\end{aligned}$$

   We now use that $\mu_1 = \mu_2 = 0$ and $\lambda_{12} = \lambda_{21}$ to get

$$\begin{aligned}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) &= x_1[\lambda_{11}x_1+\lambda_{12}x_2]+x_2[\lambda_{21}x_1+\lambda_{22}x_2]\\ &= \lambda_{11}\,x_1^2+\lambda_{22}\,x_2^2+2\lambda_{12}\,x_1\,x_2.\end{aligned}$$

   When this expression is constant, the probability density will also be constant.

6. (d) Equations (2.73) and (2.75) in the textbook can be used to find the mean and variance in the conditional density. To use these equations, we first invert Σ. Since Σ is a two-by-two matrix, this can be done by switching the diagonal elements, flipping the signs of the off-diagonal elements, and dividing by the determinant which is $|\Sigma| = (2/3)^2 - (1/3)^2 = 1/3$.

$$\Sigma^{-1} = 3 \cdot \begin{pmatrix} 2/3 & -1/3 \\ -1/3 & 2/3 \end{pmatrix} = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}.$$

Using equation (2.73), we find $\sigma^2 = 2^{-1} = 1/2$. Using equation (2.75) we find $\mu = 1 - 1/2 \cdot (-1) \cdot (y - (-2)) = 1 + 1/2\,(y + 2)$.

7. (c) The integrand in $\int \exp\left(-\frac{1}{2}(\mathbf{x}_b - \mathbf{\Lambda}_{bb}^{-1}\mathbf{m})^T \mathbf{\Lambda}_{bb}(\mathbf{x}_b - \mathbf{\Lambda}_{bb}^{-1}\mathbf{m})\right) d\mathbf{x}_b$ is the unnormalized Gaussian density. Hence this integral is the reciprocal of the normalization constant:

$$\frac{1}{(2\,\pi)^{D/2}} \frac{1}{|\mathbf{\Lambda}_{bb}^{-1}|^{1/2}} \int \exp\left(-\frac{1}{2}(\mathbf{x}_b - \mathbf{\Lambda}_{bb}^{-1}\mathbf{m})^T \mathbf{\Lambda}_{bb}(\mathbf{x}_b - \mathbf{\Lambda}_{bb}^{-1}\mathbf{m})\right) d\mathbf{x}_b = 1 \Leftrightarrow$$
$$\int \exp\left(-\frac{1}{2}(\mathbf{x}_b - \mathbf{\Lambda}_{bb}^{-1}\mathbf{m})^T \mathbf{\Lambda}_{bb}(\mathbf{x}_b - \mathbf{\Lambda}_{bb}^{-1}\mathbf{m})\right) d\mathbf{x}_b = \left(\frac{1}{(2\,\pi)^{D/2}} \frac{1}{|\mathbf{\Lambda}_{bb}^{-1}|^{1/2}}\right)^{-1}$$

8. (c)
9. (c)
10. (a)
11. (b)
12. (a)
13. (c)
14. (d)
15. (a)
16. (a) If we want to find the ML estimate of a parameter θ, we should find the value of θ for which the derivative of the likelihood, or, equivalently, the log-likelihood, of the observations is zero.Thus we want to find θ such that

$$\frac{\partial \ln(p(\{x_n | n = 1, \ldots, N\}|\theta))}{\partial\,\theta} = 0.$$

That is, we wish to find the root of

$$\frac{\partial \ln(p(\{x_n | n = 1, \ldots, N\}|\theta))}{\partial\,\theta}.$$

The Robbins-Monro procedure finds the root of a function $f(\theta)$ by updates of the form $\theta^{(N)} = \theta^{(N-1)} + a_{N-1} \cdot z(\theta^{(N-1)})$, where $z(\theta^{(N-1)})$ is the observed value of $f(\theta^{(N-1)})$. In our case, the function f is

$$\frac{\partial \ln(p(\{x_n | n = 1, \ldots, N\}|\theta))}{\partial\,\theta}.$$

For each observed x, the derivative of the log-likelihood can be calculated. This gives a value $z(\theta)$, i.e. an observation of $f(\theta)$. Hence we have

$$z(\theta^{N-1}) = \frac{\partial \ln(p(x_N|\theta^{N-1}))}{\partial \theta^{N-1}}.$$

DTU, October 2013,

Laura Frølich and Ole Winther