1. (i) Which of the following courses have you taken?

   Artificial Intelligence II Introduction to Data Mining (Advanced) Algorithms

   (ii) Have you taken any course on Probability/Statistics? If yes, please write down the course title (not number) and department.[1]

   (iii) Have you taken any course on Numerical Methods/Linear Algebra? If yes, please write down the course title (not number) and department.[1]

2. Consider a bivariate Gaussian distribution over random variables (X_1, X_2, X_3). Let $\Sigma \in \mathbb{R}^{3\times 3}$ denote the covariance matrix. Given n samples $\{(x_1^{(i)}, x_2^{(i)}, x_3^{(i)}), i = 1, \ldots, n\}$ from the distribution, how would you compute an estimate $\hat{\Sigma}_n$ of the covariance matrix Σ? Are there any assumptions you need to make on the samples for your estimate to be meaningful?

3. Let $A \in \mathbb{R}^{3\times 3}$ be a positive definite matrix and $b \in \mathbb{R}^{3\times 1}$ be a vector. How would you solve the following linear system of equations to obtain $\mathbf{x} \in \mathbb{R}^3$:

$$A\mathbf{x} = \mathbf{b} \ .$$

   It is sufficient to outline the method at a high level.

4. Assume 90% of all emails sent is spam. Based on this information, a machine learning startup Spamalot Inc. claims to have solved the email spam detection problem: for any new email, with probability 0.9 they categorizes the email as 'spam' and otherwise the email is marked as 'not spam.' Will the method correctly categorize all future emails? If yes, briefly explain why. If not, what will be the accuracy of the method? Briefly explain your answer.

CSci 5521: Fall'19
Introduction To Machine Learning
Homework 0 Solutions
(Due Monday, September 09, 11:59 pm)

2. We can use the sample covariance to estimate the population covariance Σ, where the term *population* refers to the distribution from which the sample was taken. Let us use $\mathbf{w}^t \in \mathbb{R}^3$ to denote $(x_1^{(i)}, x_2^{(i)}, x_3^{(i)})$ for $i = 1, \ldots, n$. We compute the sample covariance:

$$\hat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n} \left(\mathbf{w}^t - \bar{\mathbf{x}}\right)\left(\mathbf{w}^t - \bar{\mathbf{x}}\right)^T ,$$

   where $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}^t$ is the sample mean.

3. Since A is a positive definite matrix, it must be invertible. We can find the inverse A^{-1} using a method such as Gaussian elimination. Then

$$\begin{aligned} A\mathbf{x} &= \mathbf{b} \\ A^{-1}A\mathbf{x} &= A^{-1}\mathbf{b} \\ \mathbf{x} &= A^{-1}\mathbf{b}. \end{aligned}$$

4. No. For each email, the accuracy will be 0.82.

   Let (x, y) denote the email and the true label, where $y \in \{\text{spam}, \text{not spam}\}$. We have $\mathbb{P}\{y = \text{spam}\} = 0.9$ and $\mathbb{P}\{y = \text{not spam}\} = 0.1$. Let $\hat{y}$ be the prediction given by the startup.

   Based on Bayes rule, we have

$$\begin{aligned} \mathbb{P}\{\hat{y} = y\} &= P\{\hat{y} = \text{spam} | y = \text{spam}\}P\{y = \text{spam}\} + P\{\hat{y} = \text{not spam} | y = \text{ not spam}\}P\{y = \text{not spam}\} \\ &\overset{(a)}{=} P\{\hat{y} = \text{spam}\}P\{y = \text{spam}\} + P\{\hat{y} = \text{not spam}\}P\{y = \text{not spam}\} \\ &= 0.9 \times 0.9 + 0.1 \times 0.1 \\ &= 0.82 \end{aligned}$$

   (a): Event $\{\hat{y} = \text{spam}\}$ and event $\{y = \text{spam}\}$ are independent. Event $\{\hat{y} = \text{not spam}\}$ and event $\{y = \text{not spam}\}$ are independent as well.

1. (30 points) Consider doing least squares regression based on a training set $\mathcal{Z}_{\text{train}} = \{(x^t, r^t), t = 1, \ldots, N\}$, where $x^t \in \mathbb{R}$ and $r^t \in \mathbb{R}$.

   (i) (10 points) Consider fitting a linear model of the form

$$g_1(x) = w_1 x + w_0 \ ,$$

   with unknown parameters $w_1, w_0 \in \mathbb{R}$, which are selected so as to minimize the following empirical loss:

$$E(w_1, w_0 | \mathcal{Z}_{\text{train}}) = \frac{1}{N}\sum_{t=1}^{N} (r^t - (w_1 x^t + w_0))^2 \ .$$

   Derive the optimal values of (w_1, w_0) clearly showing all steps of the derivation.

   (ii) (10 points) Consider fitting a quadratic model of the form

$$g_2(x) = v_2 x^2 + v_1 x + v_0 \ ,$$

   with unknown parameters $v_2, v_1, v_0 \in \mathbb{R}$, which are selected so as to minimize the following empirical loss:

$$E(v_2, v_1, v_0 | \mathcal{Z}_{\text{train}}) = \frac{1}{N}\sum_{t=1}^{N} (r^t - (v_2(x^t)^2 + v_1 x^t + v_0))^2 \ .$$

   Derive the optimal values of v_2, v_1, v_0 clearly showing all steps of the derivation.

   (iii) (5 points) For a given training set $\mathcal{Z}_{\text{train}}$, let (w_1^*, w_0^*) be the optimal values of (w_1, w_0) in (i) above, and let (v_2^*, v_1^*, v_0^*) be the optimal values of (v_2, v_1, v_0) in (ii) above. Professor Gopher claims that the following is true for any given $\mathcal{Z}_{\text{train}}$:

$$E(v_2^*, v_1^*, v_0^* | \mathcal{Z}_{\text{train}}) \leq E(w_1^*, w_0^* | \mathcal{Z}_{\text{train}})$$

   Is Professor Gopher's claim correct? Clearly explain your answer.[1]

   (iv) (5 points) Consider a test set $\mathcal{Z}_{\text{test}}$ with samples which were not available during training. Assuming (w_1^*, w_0^*) and (v_2^*, v_1^*, v_0^*) are respectively the optimal parameters obtained during training using $\mathcal{Z}_{\text{train}}$ in (i) and (ii) above. Professor Gopher claims that the following is true for any test set $\mathcal{Z}_{\text{test}}$:

$$E(v_2^*, v_1^*, v_0^* | \mathcal{Z}_{\text{test}}) \leq E(w_1^*, w_0^* | \mathcal{Z}_{\text{test}})$$

   Is Professor Gopher's claim correct? Clearly explain your answer.[1]

Homework 1 Solutions

1. (i) The empirical loss function is

$$\begin{aligned} E(w_1, w_0 \mid \mathcal{Z}_{\text{train}}) &= \frac{1}{N}\sum_{t=1}^{N} (r^t - (w_1 x^t + w_0))^2 \\ &= \frac{1}{N}\sum_{t=1}^{N} ((r^t)^2 - 2w_1 x^t r^t - 2w_0 r^t + 2w_1 w_0 x^t + w_1^2 (x^t)^2 + w_0^2). \end{aligned}$$

   To find the minimum point of this function, we take the partial derivatives with respect to w_1 and w_0 and set them equal to 0:

$$\begin{aligned} \frac{\partial}{\partial w_0} = \frac{-2}{N}\sum_{t=1}^{N} r^t + \frac{2}{N}\sum_{t=1}^{N} w_1 x^t + 2w_0 = 0 \\ w_0 = \frac{1}{N}\sum_{t=1}^{N} r^t - \frac{1}{N}\sum_{t=1}^{N} w_1 x^t \\ = \bar{r} - w_1 \bar{x} \end{aligned}$$

$$\begin{aligned} \frac{\partial}{\partial w_1} = \frac{-2}{N}\sum_{t=1}^{N} x^t r^t + \frac{2}{N}\sum_{t=1}^{N} w_0 x^t + \frac{2}{N}\sum_{t=1}^{N} w_1 (x^t)^2 = 0 \\ w_1 \sum_{t=1}^{N} (x^t)^2 = \sum_{t=1}^{N} x^t r^t - \sum_{t=1}^{N} w_0 x^t \\ w_1 \sum_{t=1}^{N} (x^t)^2 = \sum_{t=1}^{N} x^t r^t - (\bar{r} - w_1 \bar{x})\sum_{t=1}^{N} x^t \\ w_1 \sum_{t=1}^{N} (x^t)^2 = \sum_{t=1}^{N} x^t r^t - N\bar{x}\bar{r} + w_1 N\bar{x}^2 \\ w_1 \sum_{t=1}^{N} (x^t)^2 - w_1 N\bar{x}^2 = \sum_{t=1}^{N} x^t r^t - N\bar{x}\bar{r} \\ w_1 = \frac{\sum_t x^t r^t - N\bar{x}\bar{r}}{\sum_t (x^t)^2 - N\bar{x}^2}. \end{aligned}$$

   (ii) Here, the empirical loss function is

$$\begin{aligned} E(v_2, v_1, v_0 \mid \mathcal{Z}_{\text{train}}) &= \frac{1}{N}\sum_{t=1}^{N} (r^t - (v_2(x^t)^2 + v_1 x^t + v_0))^2 \\ &= \frac{1}{N}\sum_{t=1}^{N} ((r^t)^2 - 2v_2(x^t)^2 r^t - 2v_1 x^t r^t - 2v_0 r^t + v_2^2 (x^t)^4 + 2v_2 v_1 (x^t)^3 \\ &\quad + v_1^2 (x^t)^2 + 2v_2 v_0 (x^t)^2 + 2v_1 v_0 x^t + v_0^2). \end{aligned}$$

   To find the minimum point of this function, we take the partial derivatives with respect to v_2, v_1, and v_0 and set them equal to 0:

$$\begin{aligned} \frac{\partial}{\partial v_2} = \frac{-2}{N}\sum_{t=1}^{N} (x^t)^2 r^t + \frac{2}{N}\sum_{t=1}^{N} v_2 (x^t)^4 + \frac{2}{N}\sum_{t=1}^{N} v_1 (x^t)^3 + \frac{2}{N}\sum_{t=1}^{N} v_0 (x^t)^2 = 0 \\ v_2 \sum_{t=1}^{N} (x^t)^4 + v_1 \sum_{t=1}^{N} (x^t)^3 + v_0 \sum_{t=1}^{N} (x^t)^2 = \sum_{t=1}^{N} (x^t)^2 r^t \end{aligned}$$

$$\begin{aligned} \frac{\partial}{\partial v_1} = \frac{-2}{N}\sum_{t=1}^{N} x^t r^t + \frac{2}{N}\sum_{t=1}^{N} v_2 (x^t)^3 + \frac{2}{N}\sum_{t=1}^{N} v_1 (x^t)^2 + \frac{2}{N}\sum_{t=1}^{N} v_0 x^t = 0 \\ v_2 \sum_{t=1}^{N} (x^t)^3 + v_1 \sum_{t=1}^{N} (x^t)^2 + v_0 \sum_{t=1}^{N} x^t = \sum_{t=1}^{N} x^t r^t \end{aligned}$$

$$\begin{aligned} \frac{\partial}{\partial v_0} = \frac{-2}{N}\sum_{t=1}^{N} r^t + \frac{2}{N}\sum_{t=1}^{N} v_2 (x^t)^2 + \frac{2}{N}\sum_{t=1}^{N} v_1 x^t + 2v_0 = 0 \\ v_2 \sum_{t=1}^{N} (x^t)^2 + v_1 \sum_{t=1}^{N} x^t + N v_0 = \sum_{t=1}^{N} r^t \end{aligned}$$

   This system of equations can be written in matrix form as

$$\begin{bmatrix} \sum_t (x^t)^4 & \sum_t (x^t)^3 & \sum_t (x^t)^2 \\ \sum_t (x^t)^3 & \sum_t (x^t)^2 & \sum_t x^t \\ \sum_t (x^t)^2 & \sum_t x^t & N \end{bmatrix} \begin{bmatrix} v_2 \\ v_1 \\ v_0 \end{bmatrix} = \begin{bmatrix} \sum_t (x^t)^2 r^t \\ \sum_t x^t r^t \\ \sum_t r^t \end{bmatrix}$$

$$\begin{bmatrix} v_2 \\ v_1 \\ v_0 \end{bmatrix} = \begin{bmatrix} \sum_t (x^t)^4 & \sum_t (x^t)^3 & \sum_t (x^t)^2 \\ \sum_t (x^t)^3 & \sum_t (x^t)^2 & \sum_t x^t \\ \sum_t (x^t)^2 & \sum_t x^t & N \end{bmatrix}^{-1} \begin{bmatrix} \sum_t (x^t)^2 r^t \\ \sum_t x^t r^t \\ \sum_t r^t \end{bmatrix}.$$

   (iii) Professor Gopher's claim is correct. To confirm this, we can note that the linear model $g_1(x) = w_1 x + w_0$ is a special case of the quadratic model $g_2(x) = v_2 x^2 + v_1 x + v_0$ where $v_2 = 0$. Since (v_2^*, v_1^*, v_0^*) are selected so as to minimize the empirical loss, we will have

$$E(v_2^*, v_1^*, v_0^* \mid \mathcal{Z}_{\text{train}}) = E(w_1^*, w_0^* \mid \mathcal{Z}_{\text{train}})$$

   if $v_2^* = 0$, and

$$E(v_2^*, v_1^*, v_0^* \mid \mathcal{Z}_{\text{train}}) < E(w_1^*, w_0^* \mid \mathcal{Z}_{\text{train}})$$

   if $v_2^* \neq 0$. In either case, therefore, it is true that

$$E(v_2^*, v_1^*, v_0^* \mid \mathcal{Z}_{\text{train}}) \leq E(w_1^*, w_0^* \mid \mathcal{Z}_{\text{train}}).$$

   (iv) Professor Gopher's claim is *not* correct. No information about $\mathcal{Z}_{\text{test}}$ was used to obtain (w_1^*, w_0^*) and (v_2^*, v_1^*, v_0^*). If, for example, the true distribution of $\mathcal{Z}$ is a linear distribution with noise

$$\beta_1 x + \beta_0 + \epsilon,$$

   it is likely that the quadratic model will learn the noise ϵ, which will cause overfitting. So it is not guaranteed that the larger model will have a lower empirical loss on the test set $\mathcal{Z}_{\text{test}}$.

2. (i) $\text{tr}(A) = 701$, $\text{tr}(A^T) = 701$, $\text{tr}(A^T A) = 484533$, $\text{tr}(AA^T) = 484533$.

```
import numpy as np
A = np.array([[1, 1, 1, 1, 1], [1, 2, 4, 8, 16], [1, 3, 9, 27, 81], [1, 4, 16, 64, 256], [1, 5, 25, 125, 625]])
print(np.trace(A))
print(np.trace(A.T) )
print(np.trace(np.dot(A.T, A)))
print(np.trace(np.dot(A, A.T)))
```

   (ii) The absolute value of $|A|$ is the volume of 5-dimensional parallelotope spanned by the column or row vectors of the matrix.

   (iii) The $|A| \neq 0$. Because the rows/columns in A are all linearly independent, thus the volume of the parallelotope spanned by the column or row vectors of A cannot be 0.

   (iv) The $\text{rank}(A) = 5$. By definition, the rank of a matrix is the dimension of the vector-space spanned by its column vectors, which is the number of linearly independent rows/columns in this matrix. A has rank 5 (full rank) because rows/columns in A are all linearly independent.

1. (30 points) Let $\mathcal{X} = \{x_1, \ldots, x_n\}$ be a set of n samples drawn i.i.d. from an univariate distribution with density function $p(x|\theta)$, where θ is an unknown parameter. In general, θ will belong to a specified subset of $\mathbb{R}$, the set of reals. For the following choices of $p(x|\theta)$, derive the maxmimum likelihood estimate of θ based on the samples $\mathcal{X}$:[1]

   (a) (7.5 points) $p(x|\theta) = \frac{1}{\sqrt{2\pi\theta}} \exp\left(-\frac{x^2}{2\theta}\right), \theta > 0$.

   (b) (7.5 points) $p(x|\theta) = \frac{1}{\theta} \exp\left(-\frac{x}{\theta}\right), 0 \leq x < \infty, \theta > 0$.

   (c) (7.5 points) $p(x|\theta) = \theta x^{\theta-1}, 0 \leq x \leq 1, 0 < \theta < \infty$.

   (d) (7.5 points) $p(x|\theta) = \frac{1}{\theta}, 0 \leq x \leq \theta, \theta > 0$.

2. (20 points) Let $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}, \mathbf{x}_i \in \mathbb{R}^d$ be a set of n samples drawn i.i.d. from a multivariate Gaussian distribution in $\mathbb{R}^d$ with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$. Recall that the density function of a multivariate Gaussian distribution is given by:

$$p(\mathbf{x}|\mu, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)\right] \ .$$

   (a) (10 points) Derive the maximum likelihood estimates for the mean μ and covariance Σ based on the sample set $\mathcal{X}$.[1,2]

   (b) (5 points) Let $\hat{\mu}_n$ be the maximum likelihood estimate of the mean. Is $\hat{\mu}_n$ a biased estimate of the true mean μ? Clearly justify your answer by computing $E[\hat{\mu}_n]$.

   (c) (5 points) Let $\hat{\Sigma}_n$ be the maximum likelihood estimate of the covariance matrix. Is $\hat{\Sigma}_n$ a biased estimate of the true covariance Σ? Clearly justify your answer by computing $E[\hat{\Sigma}_n]$.

Homework 2 Solutions

1. (a) Given n i.i.d. samples $\mathcal{X} = \{x_1, \ldots, x_n\}$, we have the log likelihood:

$$\begin{aligned} \mathcal{L}(\theta|\mathcal{X}) &= \log \prod_{i=1}^{n} p(x_i|\theta) \\ &= \log \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\theta}} \exp\left(-\frac{x_i^2}{2\theta}\right) \\ &= n \log \frac{1}{\sqrt{2\pi\theta}} - \sum_{i=1}^{n} \frac{x_i^2}{2\theta}. \end{aligned}$$

   Take the derivative of the log-likelihood function w.r.t. θ and set it to be 0:

$$\frac{\partial \mathcal{L}(\theta|\mathcal{X})}{\partial \theta} = -\frac{n}{\theta} + \frac{1}{\theta^3}\sum_{i=1}^{n} x_i^2 = 0.$$

   Solving the above equation and considering the constraint that $\theta > 0$, we have the maximum likelihood estimate of θ as:

$$\hat{\theta}^2 = \frac{\sum_{i=1}^{n} x_i^2}{n}.$$

   and

$$\hat{\theta} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}}.$$

   (b) Given n i.i.d. samples $\mathcal{X} = \{x_1, \ldots, x_n\}$, we have the log likelihood:

$$\begin{aligned} \mathcal{L}(\theta|\mathcal{X}) &= \log \prod_{i=1}^{n} p(x_i|\theta) \\ &= \log \prod_{i=1}^{n} \frac{1}{\theta} \exp\left(-\frac{x_i}{\theta}\right) \\ &= n \log \frac{1}{\theta} - \sum_{i=1}^{n} \frac{x_i}{\theta}. \end{aligned}$$

   Take the derivative of the log-likelihood function w.r.t. θ and set it to be 0:

$$\frac{\partial \mathcal{L}(\theta|\mathcal{X})}{\partial \theta} = -\frac{n}{\theta} + \frac{\sum_{i=1}^{n} x_i}{\theta^2} = 0.$$

   Solving the above equation and considering the constraint that $\theta > 0$, we have the maximum likelihood estimate of θ as:

$$\hat{\theta} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

   (c) Given n i.i.d. samples $\mathcal{X} = \{x_1, \ldots, x_n\}$, we have the log likelihood:

$$\begin{aligned} \mathcal{L}(\theta|\mathcal{X}) &= \log \prod_{i=1}^{n} p(x_i|\theta) \\ &= n \log \theta + (\theta - 1)\sum_{i=1}^{n} \log x_i \end{aligned}$$

   Take the derivative of the log-likelihood function w.r.t. θ and set it to be 0:

$$\frac{\partial \mathcal{L}(\theta|\mathcal{X})}{\partial \theta} = \frac{n}{\theta} + \sum_{i=1}^{n} \log x_i = 0.$$

   We have the maximum likelihood estimate of θ as:

$$\hat{\theta} = \frac{-n}{\sum_{i=1}^{n} \log x_i}.$$

   (d) Given n i.i.d. samples $\mathcal{X} = \{x_1, \ldots, x_n\}$, we have the log likelihood:

$$\begin{aligned} \mathcal{L}(\theta|\mathcal{X}) &= \log \prod_{i=1}^{n} p(x_i|\theta) \\ &= \log \prod_{i=1}^{n} \frac{1}{\theta} \\ &= n \log \frac{1}{\theta}, \end{aligned}$$

   where $0 \leq x_i \leq \theta, \forall x_i \in \{x_1, \ldots, x_n\}$.
   To obtain a maximum likelihood estimate of θ, we have to choose a θ that maximizes the above function and satisfies the constraint that $0 \leq x_i \leq \theta, \forall x_i \in \{x_1, \ldots, x_n\}$. $\mathcal{L}(\theta|\mathcal{X})$ is a monotonically decreasing function w.r.t θ. To maximize $\mathcal{L}(\theta|\mathcal{X})$, we choose θ be the minimal value satisfying its constraint. So we have

$$\hat{\theta} = \max_{1 \leq i \leq n} \{x_i\}.$$

2. (a) Given n i.i.d. samples $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, we have the log likelihood:

$$\begin{aligned} \mathcal{L}(\mu, \Sigma|\mathcal{X}) &= \log \prod_{i=1}^{n} p(\mathbf{x}_i|\mu, \Sigma) \\ &= \log \prod_{i=1}^{n} \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}_i - \mu)^T \Sigma^{-1} (\mathbf{x}_i - \mu)\right] \\ &= -\frac{nd}{2}\log 2\pi - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n} (\mathbf{x}_i - \mu)^T \Sigma^{-1} (\mathbf{x}_i - \mu) \end{aligned}$$

   To get the maximum likelihood estimate of μ, we take the partial derivative w.r.t. μ and set it to be 0:

$$\begin{aligned} \frac{\partial \mathcal{L}(\mu, \Sigma|\mathcal{X})}{\partial \mu} &= \frac{\partial\{-\frac{1}{2}\sum_{i=1}^{n} (\mathbf{x}_i - \mu)^T \Sigma^{-1} (\mathbf{x}_i - \mu)\}}{\partial \mu} \\ &= \sum_{i=1}^{n} \Sigma^{-1} (\mathbf{x}_i - \mu) \\ &= \mathbf{0}. \end{aligned}$$

   The second equality follows from the facts that Σ^{-1} is symmetric and $(\mathbf{x}_i - \mu)$ does not depend on Σ. To solve the above equation, we multiply both sides by Σ:

$$\sum_{i=1}^{n} (\mathbf{x}_i - \mu) = \mathbf{0}$$

   and

$$\hat{\mu}_n = \frac{\sum_{i=1}^{n} \mathbf{x}_i}{n}$$

   To get the maximum likelihood estimate of Σ, we take the partial derivative w.r.t. Σ and set it to be 0:

$$\begin{aligned} \frac{\partial \mathcal{L}(\mu, \Sigma|\mathcal{X})}{\partial \Sigma} &= \frac{\partial}{\partial \Sigma}\{-\frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n} (\mathbf{x}_i - \mu)^T \Sigma^{-1} (\mathbf{x}_i - \mu)\} \\ &= -\frac{n}{2}\cdot\frac{\partial \log(|\Sigma|)}{\partial \Sigma} - \frac{1}{2}\sum_{i=1}^{n} \frac{\partial (\mathbf{x}_i - \mu)^T \Sigma^{-1} (\mathbf{x}_i - \mu)}{\partial \Sigma} \\ &= -\frac{n}{2}\Sigma^{-T} - \frac{1}{2}\sum_{i=1}^{n} \left[-\Sigma^{-T} (\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^T \Sigma^{-T}\right] \\ &\overset{(a)}{=} -\frac{n}{2}\Sigma^{-1} - \frac{1}{2}\sum_{i=1}^{n} \left[-\Sigma^{-1} (\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^T \Sigma^{-1}\right] \\ &= \mathbf{0}, \end{aligned}$$

   where (a) is true becuase a inverse of a symmetric matrix is also symmetric. To solve the above equation, we multiply both sides by Σ to the left and get

$$n\mathbf{I} = \sum_{i=1}^{n} \left[(\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^T \Sigma^{-1}\right]$$

   Then multiply both sides by Σ to the right and get

$$n\Sigma = \sum_{i=1}^{n} (\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^T$$

$$\Sigma = \frac{\sum_{i=1}^{n} (\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^T}{n}$$

   Plugging in the maximum likelihood estimate $\hat{\mu}_n$ into the above equation, we have the maximum likelihood estimate $\hat{\Sigma}$ as

$$\hat{\Sigma}_n = \frac{\sum_{i=1}^{n} (\mathbf{x}_i - \hat{\mu}_n)(\mathbf{x}_i - \hat{\mu}_n)^T}{n}$$

   (b)

$$\begin{aligned} \mathbb{E}\left[\hat{\mu}_n\right] &= \mathbb{E}\left[\frac{\sum_{i=1}^{n} \mathbf{x}_i}{n}\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\mathbf{x}_i\right] \\ &= \frac{n\mu}{n} \\ &= \mu \end{aligned}$$

   So $\hat{\mu}_n$ is an unbiased estimate of the true mean μ.

   (c)

$$\begin{aligned} \mathbb{E}\left[\hat{\Sigma}_n\right] &= \mathbb{E}\left[\frac{\sum_{i=1}^{n} (\mathbf{x}_i - \hat{\mu}_n)(\mathbf{x}_i - \hat{\mu}_n)^T}{n}\right] \\ &= \mathbb{E}\left[\frac{\sum_{i=1}^{n}}{n}\left(\mathbf{x}_i\mathbf{x}_i^T - \mathbf{x}_i\hat{\mu}_n^T - \hat{\mu}_n\mathbf{x}_i^T + \hat{\mu}_n\hat{\mu}_n^T\right)\right] \\ &= \mathbb{E}\left[\frac{\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^T}{n} - \hat{\mu}_n\hat{\mu}_n^T - \hat{\mu}_n\hat{\mu}_n^T + \hat{\mu}_n\hat{\mu}_n^T\right] \\ &= \mathbb{E}\left[\frac{\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^T}{n} - \hat{\mu}_n\hat{\mu}_n^T\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \underbrace{\mathbb{E}\left[\mathbf{x}_i\mathbf{x}_i^T\right]}_{T_1} - \underbrace{\mathbb{E}\left[\hat{\mu}_n\hat{\mu}_n^T\right]}_{T_2} \end{aligned}$$

$$\begin{aligned} T_1 &= \mathbb{E}\left[\mathbf{x}_i\mathbf{x}_i^T\right] \\ &= \mathbb{E}\left[(\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^T\right] + \mu\mu^T \\ &= \Sigma + \mu\mu^T \end{aligned}$$

$$
\begin{aligned}
T_2 &= \mathbb{E}\left[\hat{\mu}_n \hat{\mu}_n^T\right] \\
&= \mathbb{E}\left[(\hat{\mu}_n - \mathbb{E}[\hat{\mu}_n])(\hat{\mu}_n - \mathbb{E}[\hat{\mu}_n])^T\right] + \mathbb{E}[\hat{\mu}_n]\mathbb{E}[\hat{\mu}_n]^T \\
&= Var(\hat{\mu}_n) + \mu\mu^T \\
&\overset{(a)}{=} \frac{\Sigma}{n} + \mu\mu^T,
\end{aligned}
$$

where Var means variance.
(a) is true because $\hat{\mu}_n = \frac{\sum_i^n \mathbf{x}_i}{n}$ and $Var(\hat{\mu}_n) = Var(\frac{\sum_i^n \mathbf{x}_i}{n}) = \frac{\sum_i^n Var(\mathbf{x}_i)}{n^2} = \frac{\sum_{i=1}^n \Sigma}{n^2} = \frac{\Sigma}{n}$.
Plugging T_1 and T_2 into the first equation, we have

$$
\begin{aligned}
\mathbb{E}\left[\hat{\Sigma}_n\right] &= \frac{1}{n}\sum_{i=1}^{n}(\Sigma + \mu\mu^T) - \left(\frac{\Sigma}{n} + \mu\mu^T\right) \\
&= \Sigma - \frac{\Sigma}{n} \\
&= \frac{n-1}{n}\Sigma.
\end{aligned}
$$

So $\hat{\Sigma}$ is a biased estimate of the true covariance matrix Σ.

1. **(25 points)** Let $\mathcal{X} = \{x_1, \ldots, x_n\}$ with $x_i \in \mathbb{R}$ be n samples drawn independently from a univariate Gaussian distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 \in \mathbb{R}_{++}$, i.e., $\sigma^2 > 0$.

(a) (10 points) Starting from the density of a univariate Gaussian distribution, clearly explain how to compute the maximum likelihood estimates $\hat{\mu}_n$ and $\hat{\sigma}_n^2$ for the mean μ and variance σ^2.

$P(x_1, x_2, \ldots, x_n | \mu, \sigma) = \prod_{i=1}^{n} P(x_i | \mu, \sigma)$ (+1) $\Rightarrow \log P(x_1, x_2, \ldots, x_n | \mu, \sigma) = \sum_{i=1}^{n} \log P(x_i)$ (+1)

$\max_{(\mu,\sigma)} \log P(x_1, \ldots, x_n | \mu, \sigma) = \min_{(\mu,\sigma)} -\log P(x_1, \ldots, x_n | \mu, \sigma) = \min_{(\mu,\sigma)} \sum^{n} -\log P(x_i)$ (+1) $L(\mu, \sigma)$

$= \min \sum_{i=1}^{n} -\log\left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i-\mu)}{2\sigma^2}\right)\right) \Rightarrow \min_{(\mu,\sigma)}\left(\frac{n}{2}\log 2\pi + n\log\sigma\right.$ (+1)

$\left. + \sum_{i=1}^{n} (x_i - \mu)^2/2\sigma^2\right)$

$\Rightarrow \frac{\partial L}{\partial \mu} = 0 \Rightarrow \sum_i (x_i - \mu) = 0 \Rightarrow \hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$ (+1) (+2)

$\frac{\partial L}{\partial \sigma} = 0 \Rightarrow \frac{n}{\sigma} - \frac{\sum_i (x_i - \hat{\mu}_n)^2}{\sigma^3} = 0 \Rightarrow \hat{\sigma}_n^2 = \frac{1}{n}\sum_i (x_i - \hat{\mu}_n)^2$ (+1) (+2)

(b) (7 points) Is the estimate $\hat{\mu}_n$ of μ unbiased, i.e., $E[\hat{\mu}_n] = \mu$? Clearly explain your answer.

$E[\hat{\mu}_n] = E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[x_i] = \frac{1}{n} n\mu = \mu$ (+1) (+2)

is unbiased (+1)

(c) (8 points) Is the estimate $\hat{\sigma}_n^2$ of μ unbiased, i.e., $E[\hat{\sigma}_n^2] = \sigma^2$? Clearly explain your answer.

$E[\hat{\sigma}_n^2] = E\left[\frac{1}{n}\sum_i (x_i - \hat{\mu}_n)^2\right] = \frac{1}{n}\sum_i E[(x_i - \hat{\mu}_n)^2]$

$E[(x_i - \hat{\mu}_n)^2] = E[x_i^2 - 2x_i\hat{\mu}_n + \hat{\mu}_n^2] = E[x_i^2] - 2E[x_i\hat{\mu}_n] + E[\hat{\mu}_n^2]$ (+1)

$E[x^2] - (E[x_i])^2 = \sigma^2 \Rightarrow E[x_i^2] = \sigma^2 + \mu^2$ (+1)

$E[x_i\hat{\mu}_n] = E\left[x_i \frac{1}{n}\sum_{j=1}^{n} x_j\right] = \frac{1}{n}\sum_{j=1}^{n} E[x_i x_j] = \frac{1}{n}[(n-1)\mu^2 + \mu^2 + \sigma^2]$ (+1)

$= \mu^2 + \frac{1}{n}\sigma^2$

$E[\hat{\mu}_n^2] = E\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j\right] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} E[x_i x_j] = \frac{1}{n^2}[(n^2-1)\mu^2 + n(\mu^2 + \sigma^2)] = \mu^2 + \frac{1}{n}\sigma^2$ (+1)

$\Rightarrow E[\hat{\sigma}_n^2] = \frac{1}{n}\sum_{i=1}^{n}\left[(\mu^2 + \sigma^2) - 2(\mu^2 + \frac{1}{n}\sigma^2) + (\mu^2 + \frac{1}{n}\sigma^2)\right]$

$= \frac{1}{n}\sum_{i=1}^{n}\left(\sigma^2 - \frac{1}{n}\sigma^2\right) = \frac{(n-1)}{n}\sigma^2$ (+1)

$\Rightarrow \hat{\sigma}_n$ is biased (+1)

3. **(20 points)** Suppose there are three kinds of bags of candies:

- $\frac{1}{4}$ are type h_1: 100% cherry candies,
- $\frac{1}{2}$ are type h_2: 50% cherry candies and 50% lime candies,
- $\frac{1}{4}$ are type h_3: 100% lime candies.

We have one type of bag, but we dont know which type it is. We draw a candy from the bag, and it turns out to be lime. What are the posterior probabilities $p(h_1|\text{candy} = \text{lime})$, $p(h_2|\text{candy} = \text{lime})$, $p(h_3|\text{candy} = \text{lime})$ of each type of bag? Clearly explain your answer, e.g., how you are using Bayes rule, and show your calculations.[1]

$P(h_i | \text{lime}) = \frac{P(\text{lime}|h_i) \cdot P(h_i)}{P(\text{lime})}$

$P(\text{lime}) = \sum_{i=1}^{3} P(h_i) P(\text{lime}|h_i) = \frac{1}{4} \times 0 + \frac{1}{2} \times \frac{1}{2} + \frac{1}{4} \times 1 = \frac{1}{2}$

required data derivation

$P(h_1|\text{lime}) = \frac{0 \times \frac{1}{4}}{\frac{1}{2}} = 0$

$P(h_2|\text{lime}) = \frac{\frac{1}{2} \times \frac{1}{2}}{\frac{1}{2}} = \frac{1}{2}$

$P(h_3|\text{lime}) = \frac{1 \times \frac{1}{4}}{\frac{1}{2}} = \frac{1}{2}$

4. **(25 points)** Consider a k-class multivariate parametric classification with independently drawn training data $Z = \{(\mathbf{x}^1, \mathbf{r}^1), \ldots, (\mathbf{x}^N, \mathbf{r}^N)\}$. Each data point $\mathbf{x}^t$ is d-dimensional with binary features, i.e., $x_j^t \in \{0, 1\}, j = 1, \ldots, d$. We will focus on training a Bernoulli Naive Bayes model for the dataset. In particular, for any feature vector $\mathbf{x} = [x_1 \cdots x_d]$ and any class C_i, we assume

$$
p(\mathbf{x}|C_i) = \prod_{j=1}^{d} p(x_j|C_i) .
$$

(a) (10 points) For a given class $C_i, i = 1, \ldots, k$ and a given feature $j, j = 1, \ldots, d$, clearly explain how to compute the maximum likelihood estimate $\hat{p}_{ij}$ corresponding to the Bernoulli probability $p(x_j = 1|C_i)$ using the training data.

Indices: $i = 1 \ldots k \rightarrow$ class
$t = 1 \ldots N \rightarrow$ data point
$j = 1 \ldots d \rightarrow$ feature

*Note for Likelihood, maximizing $\hat{p}_{ij}$ for class i and feature j is independent from maximizing other $\hat{p}_{ij}$ for a different class or feature because $\hat{p}_{ij}$ only affects probabilities with class i and feature j

necessary derivation

This: $L_{ij}(Z|p_{ij}) = \prod_{t=1}^{N} [p_{ij}^{x_j^t} (1-p_{ij})^{1-x_j^t}]^{r_i^t}$, where r_i^t is $\begin{cases} 1, & \text{if } r_i^t = i \\ 0, & \text{otherwise} \end{cases}$

$\log L_{ij} = \sum_{t=1}^{N} r_i^t [x_j^t \log p_{ij} + (1-x_j^t) \log(1-p_{ij})]$

Taking derivative and setting it to zero:

$\frac{\partial \log L_{ij}}{\partial p_{ij}} = \sum_{t=1}^{N} r_i^t \left(\frac{x_j^t}{p_{ij}} - \frac{1-x_j^t}{1-p_{ij}}\right) = 0$

$\Rightarrow \sum_{t=1}^{N} \frac{r_i^t x_j^t}{p_{ij}} = \sum_{t=1}^{N} \frac{r_i^t - r_i^t x_j^t}{1-p_{ij}}$

$\Rightarrow (1-p_{ij}) \sum_{t=1}^{N} r_i^t x_j^t = p_{ij} \sum_{t=1}^{N} (r_i^t - r_i^t x_j^t)$

$\Rightarrow \sum_{t=1}^{N} r_i^t x_j^t = p_{ij} \cdot \sum_{t=1}^{N} r_i^t$

$\Rightarrow \hat{p}_{ij} = \frac{\sum_{t=1}^{N} r_i^t x_j^t}{\sum_{t=1}^{N} r_i^t}$ Answer with a final solution (e.g. from slide) but without a clear derivation does not receive credit.

Name: Page 7 of 8

(b) (7 points) Given the training data, clearly explain how to compute the maximum likelihood estimates $\hat{p}(C_i)$ of the class prior probabilities $p(C_i)$ for each class $C_i, i = 1, \ldots, k$.

* Note that for priors there is an implicit constraint: $\sum_{i=1}^{k} P(C_i) = 1$ and $P(C_i) \geq 0$ so that a prior cannot be infinitely large

* Similar to (a), prior $P(C_i)$ only affects points that are in class i

necessary derivation

$L_i(Z|P(C_i)) = \prod_{t=1}^{N} P(C_i)^{r_i^t} (1-P(C_i))^{1-r_i^t}$, where $r_i^t = \begin{cases} 1, & \text{if } r_i^t = i \\ 0, & \text{otherwise} \end{cases}$

$\log L_i = \sum_{t=1}^{N} [r_i^t \log P(C_i) + (1-r_i^t) \log(1-P(C_i))]$

Taking derivative and setting to zero:

$\sum_{t=1}^{N} \frac{r_i^t}{P(C_i)} = \sum_{t=1}^{N} \frac{1-r_i^t}{1-P(C_i)}$

$\Rightarrow (1-P(C_i)) \sum_{t=1}^{N} r_i^t = P(C_i) \sum_{t=1}^{N} (1-r_i^t)$

$\Rightarrow \sum_{t=1}^{N} r_i^t = P(C_i) \cdot \sum_{t=1}^{N} 1$

$\Rightarrow \hat{P}(C_i) = \frac{\sum_{t=1}^{N} r_i^t}{N}$ Answer with a final solution (e.g., from slide) but without a clear derivation receives no credit.

(c) (8 points) Given a new test point $\mathbf{x}_{\text{test}}$ with binary features, clearly explain how to predict a class label $\hat{y}_{\text{test}}$ using the estimated $\hat{p}_{ij}$ and $\hat{p}(C_i)$.

necessary steps

$P(C_i | \mathbf{x}_{\text{test}}) \propto P(\mathbf{x}_{\text{test}} | C_i) \cdot P(C_i)$

$\Rightarrow P(C_i|\mathbf{x}_{\text{test}}) \propto \left[\prod_{j=1}^{d} \hat{p}_{ij}^{x_{\text{test}}(j)} (1-\hat{p}_{ij})^{(1-x_{\text{test}}(j))}\right] \cdot \hat{P}(C_i)$

① For each class i, compute $P(C_i|\mathbf{x}_{\text{test}})$

② class label $\hat{y}_{\text{test}} = \arg\max_i P(C_i|\mathbf{x}_{\text{test}})$

the derivation and formulas are required

A very high level answer (e.g., step ① & ②) without mathematic details receives no credit.

using $\hat{p}_{ij}$ and $\hat{P}(C_i)$

This study resource was shared via CourseHero.com