Midterm Review

Junyang Deng

November 9, 2022

1. What is the likelihood ratio $p(x \mid C_1) / p(x \mid C_2)$ in the case of Gaussian densities? (Selected from Chapter Four, exercise 5)

   SOLUTION

   The likelihood ratio is

$$\frac{p(x \mid C_1)}{p(x \mid C_2)} = \frac{\frac{1}{\sqrt{2\pi}\sigma_1} \exp\left[-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right]}{\frac{1}{\sqrt{2\pi}\sigma_2} \exp\left[-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right]}$$

   If we have $\sigma_1^2 = \sigma_2^2 = \sigma^2$, we have

$$\begin{aligned}\frac{p(x \mid C_1)}{p(x \mid C_2)} &= \exp\left[-\frac{(x-\mu_1)^2}{2\sigma^2} + \frac{(x-\mu_2)^2}{2\sigma^2}\right] \\ &= \exp\left[\frac{(\mu_1-\mu_2)}{\sigma^2}x + \frac{\mu_2^2-\mu_1^2}{2\sigma^2}\right] \\ &= \exp(wx + w_0)\end{aligned}$$

2. Let us say we have two variables x_1 and x_2 and we want to make a quadratic fit using them, namely $f(x_1, x_2) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1 x_2 + w_4 (x_1)^2 + w_5 (x_2)^2$. How can we find $w_i, i = 0, \ldots, 5$, given a sample of $X = \left\{x_1^t, x_2^t, r^t\right\}$? (Selected from Chapter Five, exercise 7)

   SOLUTION

   We write the fit as

$$f(x_1, x_2) = w_0 + w_1 z_1 + w_2 z_2 + w_3 z_3 + w_4 z_4 + w_5 z_5$$

   where $z_1 = x_1, z_2 = x_2, z_3 = x_1 x_2, z_4 = (x_1)^2$, and $z_5 = (x_2)^2$. We can then use linear regression to learn $w_i, i = 0, \ldots, 5$. The linear fit in the five-dimensional $(z_1, z_2, z_3, z_4, z_5)$ space corresponds to a quadratic fit in the two-dimensional (x_1, x_2) space.

3. Describe Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA), and compare the differences.

   SOLUTION

   - **PCA**: The principle of PCA is to find a mapping W from the inputs x with d-dimension to a new k-dimesional space. The projection on the direction w is then $z = W^T x$. The criterion to select W is to maximize the variance of the projected data.
     In order to do this, we need to first (1) compute the covariance matrix of x by $\Sigma = E\left[(x-\mu)(x-\mu)^T\right]$. Then, we can (2) compute eigenvalues and eigenvectors of Σ and (3) rank eigenvalues from largest to smallest. Eigenvectors that correspond to highest eigenvalues will be (4) selected as

principal components.

In this way, we found directions that maximize the variance among data.

- **LDA**: The principle of LDA is to find a direction w which separates different classes as much as possible. Criteria of selecting w is to maximize between-class variance and minimize within-class variance.

  In order to this, we should (1) construct the within-class ($\mathbf{S}_W$) and between-class ($\mathbf{S}_B$) scatter matrix by $\mathbf{S}_W = \sum_{i=1}^{K} \boldsymbol{S}_i$, $\mathbf{S}_i = \sum_t r_i^t \left(\boldsymbol{x}^t - \boldsymbol{m}_i\right)\left(\boldsymbol{x}^t - \boldsymbol{m}_i\right)^T$ and $\mathbf{S}_B = \sum_{i=1}^{K} N_i \left(\boldsymbol{m}_i - \boldsymbol{m}\right)\left(\boldsymbol{m}_i - \boldsymbol{m}\right)^T$. Based on $\mathbf{S}_W$ and $\mathbf{S}_B$, construct the Fisher's discriminant $J(W) = \dfrac{|W^T \mathbf{S}_B W|}{|W^T \mathbf{S}_W W|}$.
  (2) Then, find a W which maximizes the Fisher's discriminant. In the two-class case, the solution is $c\boldsymbol{S}_W^{-1}(\boldsymbol{m}_1 - \boldsymbol{m}_2)$.

- **Differences**

  (a) LDA is a supervised learning method, which means labels should also be included in the input, while PCA is an unsupervised learning method.

  (b) For LDA, since it has label information, it can minimize the within-class variance as well as maximize the between class variance. For PCA, it can only maximize the overall variance.

  (c) LDA has a better performance than PCA when there is more noise in the data. PCA is more sensitive to outliers.

4. To evaluate the predictive performance of a model constructed from a two-class data set, k-fold cross-validations are frequently applied. Describe the concept of cross-validation, and two performance measures, sensitivity, and specificity, respectively.

   SOLUTION

   - **Cross-validation**: Also named k-fold cross-validation. People will randomly partition the data into k mutually exclusive subsets, each of them has approximately equal size. At i-th iteration, class D_i will be used as validation set, while the rest will be used as training set. After each iteration, performance (or error) of the model will be calculated.
   - **Sensitivity**: Sensitivity is the true positive recognition rate. It can be calculated as $TP/(TP+FN)$.
   - **Specificity**: Specificity is the true negative recognition rate. It can be calculated as $TN/(TN+FP)$.

5. To learn predictive models from a data set, which is a sample from real-world data. Describe the relationship between prediction errors, bias, and variance.

   SOLUTION

   We usually use Mean Square Error (MSE) to describe prediction error. Denote the estimated parameter as d, the expectation of estimated parameter as Ed, and the underlying unknown parameter as θ.

$$\begin{aligned}
\text{MSE} &= E\left[(d-\theta)^2\right] \\
&= E\left[(d-Ed+Ed-\theta)^2\right] \\
&= E\left[(d-Ed)^2 + 2(d-Ed)(Ed-\theta) + (Ed-\theta)^2\right] \\
&= E(d-Ed)^2 + (Ed-\theta)^2
\end{aligned}$$

   $E(d-Ed)^2$ is variance and $(Ed-\theta)^2$ is bias^2. Therefore, preidction errors (or MSE) can be decompose into squared bias and variance.

6. What is the assumption of a multi-variate and supervised data set to which naive Bayesian classifier can be applied? Address your points in the aspect of variable dependency, co-variance matrix, and data distributions of the given classes.

   SOLUTION

   When variables can be assumed as independent, different given classes share a same covariance matrix which off-diagonal terms are zero, and each data point is independent, identical distributed, native Bayesian classifier can be applied.

   - Variable dependency: the naïve Bayesian classifier assume all variables are conditionally independent. They are no dependency between attributes.
   - Co-variance matrix: the naive Bayesian classifier assumes all off-diagonal terms are 0.
   - Distribution: the naïve Bayesian classifier requires an iid distribution and a Gaussian distribution is usually assumed.

7. In a two-class problem, the likelihood ratio is $p(x \mid C_1) / p(x \mid C_2)$. Write the discriminant function in terms of the likelihood ratio. (Selected from Chapter Three, exercise 2)

   SOLUTION

   We define a discriminant function as

   $$g(x) = \log \frac{P(C_1 \mid x)}{P(C_2 \mid x)} \text{ and choose } \begin{cases} C_1 & \text{if } g(x) > 0 \\ C_2 & \text{otherwise} \end{cases}$$

   Log odds is the sum of log likelihood ratio and log of prior ratio:

   $$g(x) = \log \frac{p(x \mid C_1)}{p(x \mid C_2)} + \log \frac{P(C_1)}{P(C_2)}$$

   If the priors are equal, the discriminant is the log likelihood ratio.

8. In a two-class, two-action problem, if the lost function is $\lambda_{11} = \lambda_{22} = 0, \lambda_{12} = 10$, and $\lambda_{21} = 5$, write the optimal decision rule. How does the rule change if we add a third action of reject with $\lambda_1 = 1$? (Selected from Chapter Three, exercise 4)

   SOLUTION

   We calculate the expected risks of the two actions:

   $$R(\alpha_1 \mid x) = \lambda_{11} P(C_1 \mid x) + \lambda_{12} P(C_2 \mid x) = 10 P(C_2 \mid x)$$
   $$R(\alpha_2 \mid x) = \lambda_{21} P(C_1 \mid x) + \lambda_{22} P(C_2 \mid x) = 5 P(C_1 \mid x)$$

   When reject is not considered, we choose C_1 if

   $$R(\alpha_1 \mid x) < R(\alpha_2 \mid x) \Rightarrow 10 - 10P(C_1|x) > 5P(C_1|x)$$

   Solve the inequality, we get $P(C_1|x) > 2/3$. Therefore, when $P(C_1|x) > 2/3$, choose C_1, when $P(C_1|x) < 2/3$, choose C_2.

   When reject is considered, we choose C_1 when

   $$R(\alpha_1|x) < R(\alpha_2|x) \text{ and } R(\alpha_1|x) < \lambda$$

Solve that $P(C_1|x) > 9/10$.

In summary,

$$\text{We choose} \begin{cases} C_1 & P(C_1|x) > 9/10 \\ \text{reject} & 2/3 < P(C_1|x) < 9/10 \\ C_2 & P(C_1|x) < 2/3 \end{cases}$$

9. Show equation 5.11 (Selected from Chapter Five, exercise 1)

$$p(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left(z_1^2 - 2\rho z_1 z_2 + z_2^2\right)\right]$$

SOLUTION

Given that

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

we have

$$|\Sigma| = \sigma_1^2\sigma_2^2 - \rho^2\sigma_1^2\sigma_2^2 = \sigma_1^2\sigma_2^2(1-\rho^2)$$

$$|\Sigma|^{1/2} = \sigma_1\sigma_2\sqrt{1-\rho^2}$$

$$\Sigma^{-1} = \frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)} \begin{bmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{bmatrix}$$

and $(\boldsymbol{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$ can be expanded as

$$\begin{aligned} &\begin{bmatrix} x_1-\mu_1 & x_2-\mu_2 \end{bmatrix} \begin{bmatrix} \frac{\sigma_2^2}{\sigma_1^2\sigma_2(1-\rho^2)} & -\frac{\rho\sigma_1\sigma_2}{\sigma_1^2\sigma_2(1-\rho^2)} \\ -\frac{\rho\sigma_1\sigma_2}{\sigma_1^2\sigma_2(1-\rho^2)} & \frac{\sigma_1^2}{\sigma_1^2\sigma_2(1-\rho^2)} \end{bmatrix} \begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{bmatrix} \\ &= \frac{1}{1-\rho^2}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right] \end{aligned}$$

10. Compare EM and K-Means

- Commonalities
  (a) EM and K-means start by randomly assigning K models.
  (b) EM and K-means use a same 2-step iterative approach to optimize the objective function. They optimize a while keeping b unchanged, then optimize b while keeping a unchanged.
  (c) EM and K-means can only find the local minima or maxima for their target function. The results are sensitive to initiation.
- Differences
  (a) K-means is a simplified version of EM. K-means uses a distance-based approach to assign classes. EM uses a density-based (probalistic) approach.
  (b) K-means can use different distance metrics. EM implicitly relies on the Mahalanobis distance function which is part of its density estimation approach.
  (c) K-means minimizes the squared distance of a sample to its cluster prototype. EM maximizes the log likelihood of a sample given a model; models are assumed to be mixtures of K Gaussian and their priors.

(d) The model of K-means is K centroid points. The model of EM is K priors, means and covariance matrix (!!).

(e) K-means is a hard clustering method. The result is $\{0, 1\}$. EM is a soft clustering method. The result is $[0,1]$.

(f) EM directly deals with dependencies between attributes in its density-based approach: the probability of a sample belongs to one class equals the product between class prior and the Mahalanobis distance between sample and mean. therefore, EM clusters do not depend on units of measurements and orientation of attributes in space.

11. How to choose K in K-means clustering and EM?

- Based on the application, e.g. image quantization
- Use PCA to aid this process. We can first use PCA to project data to the direction with largest variance. Then count the data points and figure out how many peaks exist. The number of peaks can be the optimal K.
- Plot the reconstruction error against K. Choose the elbow.