STAT 3690 Lecture 31

zhiyanggeezhou.github.io

Zhiyang Zhou (zhiyang.zhou@umanitoba.ca)

Apr 13, 2022

Classification

- Predictive task in which the response takes values across K discrete categories (i.e., not continuous)
  - For one subject, to predict its class label Y when its features $\mathbf{X}$ is observed
  - Binary classification: $K = 2$
  - Having training data with known class labels
  - E.g.
    * Given scanned handwritten digits: 28×28 grid of pixels each reflecting the value of grey scale; see Lec 23. From vectorized pictures determine what digit was written.
    * Predicting the region of Italy in which a brand of olive oil was made, based on its chemical composition; see Lec 29.

- Bayes classifier
  - Classify according to posterior $\Pr(Y = k \mid \mathbf{X} = \boldsymbol{x}) = f_k(\boldsymbol{x})\pi_k / \sum_{\ell=1}^{K} f_\ell(\boldsymbol{x})\pi_\ell$, $k = 1, \ldots, K$
    * $f_k(\boldsymbol{x})$: the probability density/mass function of $\mathbf{X}$ conditioning on Class k
    * $\pi_k = \Pr(Y = k)$: prior probability of Class k
  - Bayes classifier

$$h(\boldsymbol{x}) = \arg\max_{k=1,\ldots,K} \Pr(Y = k \mid \mathbf{X} = \boldsymbol{x}) = \arg\max_{k=1,\ldots,K} f_k(\boldsymbol{x})\pi_k$$

Linear discriminant analysis (LDA)

- Assuming $f_k(\boldsymbol{x})$ = density of $MVN_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$
- LDA classifier

$$h(\boldsymbol{x}) = \arg\max_{k=1,\ldots,K} \delta_k(\boldsymbol{x})$$

  - Discriminant functions $\delta_k(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \ln \pi_k$
    * Linear functions with respect to $\boldsymbol{x}$
- Empirical version
  - Training data: $\boldsymbol{x}_i \in \mathbb{R}^p$ and $y_i \in \{1, \ldots, K\}$, $i = 1, \ldots, n$
    * n_k: the number of training observations in class k, $k = 1, \ldots, K$
  - Estimation for $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}$ and π_k
    * $\hat{\pi}_k = n_k/n$
    * $\hat{\boldsymbol{\mu}}_k = n_k^{-1} \sum_{i=1}^{n} \boldsymbol{x}_i \cdot \mathbf{1}(y_i = k)$
    * $\widehat{\boldsymbol{\Sigma}} = (n-K)^{-1} \sum_{k=1}^{K} \sum_{i=1}^{n} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k)^\top \cdot \mathbf{1}(y_i = k)$
  - Empirical LDA classifier

$$\hat{h}(\boldsymbol{x}) = \arg\max_{k=1,\ldots,K} \hat{\delta}_k(\boldsymbol{x})$$

    - $\hat{\delta}_k(\boldsymbol{x}) = \boldsymbol{x}^\top \widehat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}_k - \frac{1}{2} \hat{\boldsymbol{\mu}}_k^\top \widehat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}_k + \ln \hat{\pi}_k$
- Example (Fisher's or Anderson's `iris` data)
  - 50 flowers from each of 3 species of iris: setosa, versicolor, and virginica.
  - Measurements in centimeters of the variables sepal length and width and petal length and width.

Quadratic discriminant analysis (QDA)

- Assuming $f_k(\boldsymbol{x})$ = density of $MVN_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$
- QDA classifier

$$h(\boldsymbol{x}) = \arg \max_{k=1,\dots,K} \delta_k(\boldsymbol{x})$$

  - Discriminant functions $\delta_k(\boldsymbol{x}) = -\boldsymbol{x}^\top \boldsymbol{\Sigma}_k^{-1} \boldsymbol{x} + 2\boldsymbol{x}^\top \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k - \boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k + 2 \ln \pi_k - \ln \det \boldsymbol{\Sigma}_k$
    - Quadratic functions with respect to $\boldsymbol{x}$

- Empirical version
  - Training data: $\boldsymbol{x}_i \in \mathbb{R}^p$ and $y_i \in \{1, \dots, K\}$, $i = 1, \dots, n$
    - n_k: the number of training observations in class k, $k = 1, \dots, K$
  - Estimation for $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}$ and π_k
    - $\hat{\pi}_k = n_k/n$
    - $\hat{\boldsymbol{\mu}}_k = n_k^{-1} \sum_{i=1}^n \boldsymbol{x}_i \cdot \mathbf{1}(y_i = k)$
    - $\widehat{\boldsymbol{\Sigma}}_k = (n_k - 1)^{-1} \sum_{i=1}^n (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k)^\top \cdot \mathbf{1}(y_i = k)$
  - Empirical classifier

$$\hat{h}(\boldsymbol{x}) = \arg \max_{k=1,\dots,K} \hat{\delta}_k(\boldsymbol{x})$$

    - $\hat{\delta}_k(\boldsymbol{x}) = -\boldsymbol{x}^\top \widehat{\boldsymbol{\Sigma}}_k^{-1} \boldsymbol{x} + 2\boldsymbol{x}^\top \widehat{\boldsymbol{\Sigma}}_k^{-1} \hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_k^\top \widehat{\boldsymbol{\Sigma}}_k^{-1} \hat{\boldsymbol{\mu}}_k + 2 \ln \hat{\pi}_k - \ln \det \widehat{\boldsymbol{\Sigma}}_k$
- Example (`iris` data, con'd)

Assessment

- Misclassification rate
  - Population: $\mathrm{err} = \Pr(Y \neq h(\mathbf{X}))$
  - Empirical: $\widehat{\mathrm{err}} = (n^*)^{-1} \sum_{i=1}^{n^*} \mathbf{1}\{y_i \neq \hat{h}(\boldsymbol{x}_i)\}$
    - Testing data: $\boldsymbol{x}_i \in \mathbb{R}^p$ and $y_i \in \{1, \dots, K\}$, $i = 1, \dots, n^*$
- Cross validation (CV)
  - Leave-one-out CV
  - M-fold CV
    - Leave-one-out $\Leftrightarrow$ n-fold