Machine Learning, Spring 2019
EN 2105
MWF 9-9:50am

Chao Lan
EN 4084A
clan@uwyo.edu

Gaussian Discriminant Analysis

Recall the bayes decision rule is

$$p(y = k \mid x) \geq p(y = c \mid x), \forall k \neq c. \tag{1}$$

Gaussian Discriminant Analysis first applies the Bayes' rule and obtains an equivalence

$$\frac{p(x \mid y = k) \cdot p(y = k)}{p(x)} \geq \frac{p(x \mid y = c) \cdot p(y = c)}{p(x)}, \ \forall k \neq c, \tag{2}$$

which is further equivalent to

$$p(x \mid y = k)p(y = k) \geq p(x \mid y = c)p(y = c), \ \forall k \neq c. \tag{3}$$

Then, GDA estimates $p(x \mid y = k)$ and $p(y = k)$ from data. There are different ways to construct and estimate both probabilities. We will introduce a representative one.

Part 1: $p(y = k)$. Let n be the number of training instances and n_k the number of training instances from class k. We can estimate $p(y = k)$ by

$$p(y = k) = \frac{n_k}{n}. \tag{4}$$

Part 2: $p(x \mid y = k)$. GDA assumes each class is generated from a normal distribution. With different assumptions on the distributions, GDA is further divided into QDA and LDA.

Quadratic Discriminant Analysis (QDA) assumes each class has its own covariance, i.e.,

$$p(x \mid y = k) = \mathcal{N}(\mu_k, \Sigma_k). \tag{5}$$

Linear Discriminant Analysis (LDA) assumes all classes share the same covariance, i.e.,

$$p(x \mid y = k) = \mathcal{N}(\mu_k, \Sigma). \tag{6}$$

[*Discussion*] How to interpret the different assumptions of LDA and QDA?

[*Discussion*] Which model has larger model complexity?

For both QDA and LDA, the mean μ_k can be estimated using MLE.

$$\hat{\mu}_k = \frac{1}{n_k} \sum\nolimits_{y_i = k} x_i. \tag{7}$$

For QDA, the covariance Σ_k is estimated as

$$\hat{\Sigma}_k = \frac{1}{n_k - 1} \sum\nolimits_{y_i = k} (x_i - \mu_k)(x_i - \mu_k)^T, \tag{8}$$

and for LDA, the covariance Σ is estimated as

$$\hat{\Sigma} = \frac{1}{n - K} \sum\nolimits_{k=1}^{K} \sum\nolimits_{y_i = k} (x_i - \mu_k)(x_i - \mu_k)^T. \tag{9}$$

Once $p(x \mid y = k)$ and $p(y = k)$ are estimated, GDA make prediction for a testing instance z by computing $p(z \mid y = k)$ and applying the (equivalent) bayes decision rule (3).

Naive Bayes

The Naive Bayes classifier is similar to GDA. The difference is in the construction of $p(x \mid y)$. NB makes an (naive but strong) assumption that features are independent so that

$$\begin{aligned} p(x \mid y = k) &= p(x_{\cdot 1} \mid y = k) \cdot p(x_{\cdot 2} \mid y = k) \ldots \cdot p(x_{\cdot p} \mid y = k) \\ &= \Pi_{j=1}^{p} p(x_{\cdot j} \mid y = k). \end{aligned} \tag{10}$$

There are many ways to design $p(x_{\cdot j} \mid y = k)$.

For continuous $x_{\cdot j}$ the Gaussian naive Bayes assumes

$$p(x_{\cdot j} \mid y = k) = \mathcal{N}(\mu_j, \sigma_j^2). \tag{11}$$

For discrete $x_{\cdot j}$, the Bernoulli naive Bayes assumes

$$p(x_{\cdot j} \mid y = k) = B(\theta). \tag{12}$$