36-700: Probability and Mathematical Statistics I **Fall 2016**

Lecture 32: November 28

Lecturer: Siva Balakrishnan

32.1 Review and Outline

In the last class we discussed directed graphical models.

1. Conditioning on colliders
2. d-separation

In this lecture we will discuss classification. This is Chapter 22 of the Wasserman book.

In our past lectures we have focused on regression (linear and non-parametric), and density estimation (parametric and non-parametric). A closely related task to regression is that of classification. Formally, we observe i.i.d. data $(X_1, Y_1), \ldots, (X_n, Y_n)$ where $X_i \in \mathbb{R}^d$ and $Y_i \in \{1, \ldots, k\}$, i.e. there are k classes.

A classifier or a classification rule is simply a map $h : \mathbb{R}^d \mapsto \{1, \ldots, k\}$, i.e. when we observe a new X we predict its category/class to be $h(X)$.

32.2 Error rates and binary classification

Broadly, the goal in classification is to find classification rules that are accurate. The true (population) error rate of a classifier h is:

$$L(h) = \mathbb{P}(h(X) \neq Y),$$

and the empirical error rate is:

$$\widehat{L}(h) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(h(X_i) \neq Y_i).$$

A special case of classification is when we have binary outcomes, i.e. $Y \in \{0, 1\}$. Let

$$r(x) = \mathbb{E}[Y|X = x] = \mathbb{P}(Y = 1|X = x),$$

be the usual regression function. We can re-write this using Bayes' rule:

$$r(x) = \frac{f(x|Y=1)P(Y=1)}{f(x|Y=1)P(Y=1) + f(x|Y=0)P(Y=0)}.$$

We often denote $P(Y=1) = \pi$ and then we have $P(Y=0) = 1-\pi$.

The *optimal* classifier is known as the Bayes' classifier. It is given by:

$$h^*(x) = \begin{cases} 1 & \text{if } \ r(x) > \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases}$$

For a classifier, we can define its *decision boundary*. It is generally defined as a surface that partitions the domain of X into two sets, one for each class.

The Bayes classifier has a decision boundary given by:

$$D(h^*) = \{x : P(Y=1|X=x) = P(Y=0|X=x)\}.$$

There are two other equivalent forms of the Bayes' classifier:

1.
$$h^*(x) = \begin{cases} 1 & \text{if } \ \pi f(x|Y=1) > (1-\pi) f(x|Y=0) \\ 0 & \text{otherwise.} \end{cases}$$

2.
$$h^*(x) = \begin{cases} 1 & \text{if } \ P(Y=1|X=x) > P(Y=0|X=x) \\ 0 & \text{otherwise.} \end{cases}$$

Finally, to re-visit the optimality of the Bayes rule: it is the case that for any other classification rule h, we have that:

$$L(h^*) \leq L(h),$$

so the Bayes classifier minimizes the true error rate amongst all classifiers.

The main issue however is that the Bayes classifier depends on unknown quantities, i.e. the probabilities $P(Y=1|X=x)$ or the densities $f(x|Y=1)$ and so on. However, it does serve as a template to develop classifiers. Many classifiers explicitly try to approximate the Bayes rule using the training data.

Broadly, there are different strategies for classification:

1. Empirical Risk Minimization: Here the idea is simple, we choose a set of classifiers $\mathcal{H}$ and try to find $h \in \mathcal{H}$ that minimizes some estimate of $L(h)$. Usually we use the empirical risk $\widehat{L}(h)$.

2. Regression: We estimate the regression function or $P(Y = 1|X = x)$ and then define the classfier:

$$h(x) = \begin{cases} 1 & \text{if } \widehat{P}(Y = 1|X = x) > 1/2 \\ 0 & \text{otherwise.} \end{cases}$$

3. Density Estimation: We estimate π, $f(x|Y = 0)$ and $f(x|Y = 1)$ using the training data, and then use the classifier:

$$h^*(x) = \begin{cases} 1 & \text{if } \widehat{\pi}\widehat{f}(x|Y = 1) > (1 - \widehat{\pi})\widehat{f}(x|Y = 0) \\ 0 & \text{otherwise.} \end{cases}$$

We will consider some of these ideas today and then continue in the next lecture.

32.3 Linear Discriminant Analysis

Our first classifier to consider, will be based on density estimation. First, let us hypothesize that:

$$\begin{aligned} f(x|Y = 0) &\sim N(\mu_0, \Sigma) \\ f(x|Y = 1) &\sim N(\mu_1, \Sigma) \\ P(Y = 1) &= \pi_1. \end{aligned}$$

In this simplified setting we can derive the form of the Bayes classifier. In particular, $h^*(x) = 1$ if:

$$\pi_1 \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)}{2}\right) > (1-\pi_1)\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)}{2}\right)$$

rearranging this we obtain that $h^*(x) = 1$ if,

$$\log(\pi_1/(1-\pi_1)) - \frac{(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)}{2} > -\frac{(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)}{2}.$$

We note that the decision boundary of this classifier is:

$$\log(\pi_1/(1-\pi_1)) - \frac{(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)}{2} = -\frac{(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)}{2},$$

which on re-arrangement gives:

$$\log(\pi_1/(1-\pi_1)) - \frac{\mu_1^T\Sigma^{-1}\mu_1 + \mu_0^T\Sigma^{-1}\mu_0}{2} + x^T\Sigma^{-1}(\mu_1 - \mu_0) = 0,$$

which shows that the decision boundary of the classifier is linear, i.e. of the form $\alpha_0 + \alpha^T x = 0$, for some values α_0 and α. This is why the classifier is called *linear* discriminant analysis.

We could have also, considered a setting where:

$$\begin{aligned} f(x|Y=0) &\sim N(\mu_0, \Sigma_0) \\ f(x|Y=1) &\sim N(\mu_1, \Sigma_1) \\ P(Y=1) &= \pi_1. \end{aligned}$$

Under this setting the Bayes classifier will be a quadratic function of x and this is known as Quadratic Discriminant Analysis.

Going back to LDA, now that we have a form for the Bayes classifier, we can approximate the Bayes rule by estimating the various unknown quantities. Concretely, given a training data set $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ we can estimate:

$$\begin{aligned} \widehat{\pi}_1 &= \frac{1}{n}\sum_{i=1}^n \mathbb{I}(Y_i = 1) \\ \widehat{\mu}_0 &= \frac{1}{\sum_{i=1}^n \mathbb{I}(Y_i=0)}\sum_{i=1}^n X_i \mathbb{I}(Y_i = 0) \\ \widehat{\mu}_1 &= \frac{1}{\sum_{i=1}^n \mathbb{I}(Y_i=1)}\sum_{i=1}^n X_i \mathbb{I}(Y_i = 1). \end{aligned}$$

These are the maximum likelihood estimators for these parameters. The MLE for Σ is given by:

$$\begin{aligned} \widehat{\Sigma}_0 &= \frac{1}{\sum_{i=1}^n \mathbb{I}(Y_i=0)}\sum_{i=1}^n (X_i - \widehat{\mu}_0)(X_i - \widehat{\mu}_0)^T \mathbb{I}(Y_i = 0) \\ \widehat{\Sigma}_1 &= \frac{1}{\sum_{i=1}^n \mathbb{I}(Y_i=1)}\sum_{i=1}^n (X_i - \widehat{\mu}_1)(X_i - \widehat{\mu}_1)^T \mathbb{I}(Y_i = 1) \\ \widehat{\Sigma} &= \frac{\sum_{i=1}^n \mathbb{I}(Y_i=0)\widehat{\Sigma}_0 + \sum_{i=1}^n \mathbb{I}(Y_i=1)\widehat{\Sigma}_1}{n}. \end{aligned}$$

With these estimates in place we just use the rule $h(x) = 1$ if:

$$\log(\widehat{\pi}_1/(1-\widehat{\pi}_1)) - \frac{\widehat{\mu}_1^T\widehat{\Sigma}^{-1}\widehat{\mu}_1 + \widehat{\mu}_0^T\widehat{\Sigma}^{-1}\widehat{\mu}_0}{2} + x^T\widehat{\Sigma}^{-1}(\widehat{\mu}_1 - \widehat{\mu}_0) > 0,$$

32.4 Logistic Regression

A popular direct regression based classifier is a logistic regressor. Here the hypothesis is that:

$$P(Y = 1|X = x) = \frac{\exp(\beta_0 + \beta^T x)}{1 + \exp(\beta_0 + \beta^T x)}.$$

This is a logistic function of $\beta_0 + \beta^T x$ and has the property that it is always between $[0, 1]$ and so represents a true probability. Notice the following properties, $P(Y = 1|X = x) \to 1$ if $\beta_0 + \beta^T x \to \infty$, $P(Y = 1|X = x) \to 0$ if $\beta_0 + \beta^T x \to -\infty$ and $P(Y = 1|X = x) = 1/2$ if $\beta_0 + \beta^T x = 0$

Under the logistic hypothesis we can again derive the Bayes rule, $h^*(x)$ is 1 if:

$$\frac{\exp(\beta_0 + \beta^T x)}{1 + \exp(\beta_0 + \beta^T x)} > \frac{1}{1 + \exp(\beta_0 + \beta^T x)},$$

which on rearrangement gives:

$$\beta_0 + \beta^T x > 0.$$

The decision boundary for the Bayes classifier is then simply:

$$\beta_0 + \beta^T x = 0,$$

which is again a linear decision boundary. So both LDA and logistic regression are linear classifiers. In our next class, we will discuss how to fit a logistic regression, i.e. estimate β_0 and β_1, and then compare logistic regression to LDA.