Linear Methods for Classification

Thomas Lonon

Division of Financial Engineering
Stevens Institute of Technology

February 15, 2017

The *Classification Problem* is the problem of assigning an observation to a particular class or category.

We seek to estimate f on the basis of training observations where $y_1, \ldots, y_n$ are *qualitative*

Accuracy is usually quantified using the error rate r_e given by:

$$r_e = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{y_i \neq \hat{y}_i\}}$$

We begin looking at situations where we have categorical or qualitative data types.

Rather than modeling the variables, though, we model the probabilities.

Assume that y is a binary $(0, 1)$ variable, we define:

$$p(x) = \mathbb{P}(y = 1|x)$$

How should this probability be modeled?

Assuming $\mathcal{G}$ has K classes, we can create a vector of indicator variables $Y_k, k \in 1 \ldots K$. We then take the N training instances to form the $N \times K$ *indicator response matrix* $\mathbf{Y}$. We then fit a linear regression model to the columns of $\mathbf{Y}$. This fit will be given by:

$$\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

[2]

We can see from this (Section 3.2.4) that the matrix, $\hat{\mathbf{B}}$ given by:

$$\hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

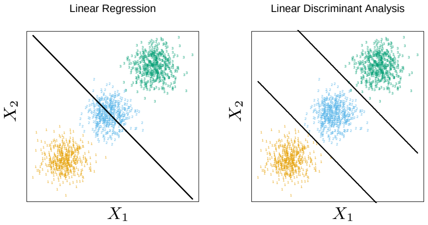
Will be our estimates for the linear regression.

We then classify a new observation (x) in the following manner:

- Compute the fitted output $\hat{f}(x)^T = (1, x^T)\hat{\mathbf{B}}$
- Identify the largest component and classify accordingly

$$\hat{G}(x) = \arg\max_{k \in \mathcal{G}} \hat{f}_k(x)$$

[2]



[2]

FIGURE 4.2. *The data come from three classes in* $\mathbb{R}^2$ *and are easily separated by linear decision boundaries. The right plot shows the boundaries found by linear discriminant analysis. The left plot shows the boundaries found by linear regression of the indicator response variables. The middle class is completely masked (never dominates).*

A function that is guaranteed to take values between 0 and 1 is the *logistic function* given by:

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

This can be rearranged into the form of the *odds ratio*:

$$\frac{p(x)}{1 - p(x)} = e^{\beta_0+\beta_1 x}$$

or we can take the logarithm of both sides to get the *logit* function:

$$\log\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 x$$

Bayes Classifier

We assign observations to a class to which it is most likely to belong: Mathematically we represent this as for observation x_0, we assign it to class k if:

$$\mathbb{P}(Y = k|X = x_0) \geq \mathbb{P}(Y = j|X = x_0), \forall j$$

The result of this classification will give you boundaries between observations known as *Bayes Decision Boundaries*

Let $\pi_k = \mathbb{P}(Y = k)$ be the *prior* probability that a randomly chosen observation comes from the k^{th} class

Let $f_k(x) = \mathbb{P}(X = x | Y = k)$ be the density of X for an observation that comes from the k^{th} class (likelihood)

We have (assuming that membership can be held in only one class):

$$\mathbb{P}(X = x) = \sum_{k=1}^{K} \mathbb{P}(X = x | Y = k)\mathbb{P}(Y = k) = \sum_{k=1}^{K} f_k(x)\pi_k$$

[1]

From probability we know that:

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}$$

or

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

which can be combined to get:

$$\mathbb{P}(B|A)\mathbb{P}(A) = \mathbb{P}(A|B)\mathbb{P}(B)$$

Using Bayes' Theorem:

$$\mathbb{P}(Y = k|X = x)\mathbb{P}(X = x) = \mathbb{P}(X = x|Y = k)\mathbb{P}(Y = k)$$

we have:

$$\mathbb{P}(Y = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{m=1}^{K} f_m(x)\pi_m}$$

and we define $p_k(x)$ given by:

$$p_k(x) = \mathbb{P}(Y = k|X = x)$$

as the probability that the observation belongs to the k^{th} class, given that its observed value is x. This is called the *posterior* probability

Let our classification function $G(x)$ produce a discrete set of responses $\mathcal{G}$, for example apples, oranges, pears, etc.

We can always divide the input space into a collection or regions according to this classification

For a class of procedures, these decision boundaries are linear, for example linear discriminant analysis

Suppose there are K classes, and the fitted linear model for the k^{th} indicator response variable is

$$\hat{f}_k(x) = \hat{\beta}_{k0} + \hat{\beta}_k^T x$$

The decision boundary between class k and ℓ is the set of points for which $\hat{f}_k(x) = \hat{f}_\ell(x)$, or the set:

$$\{x : (\hat{\beta}_{k0} - \hat{\beta}_{\ell 0}) + (\hat{\beta}_k - \hat{\beta}_\ell)^T x = 0\}$$

this is an affine set (or hyperplane although one that will potentially not pass through the origin)[2]

hyperplane: a subspace of one dimension less than its ambient space

For two classes, $G \in \mathcal{G} = \{1, 2\}$, as discussed a popular method for the posterior probabilities is

$$\mathbb{P}(G = 1|X = x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$
$$\mathbb{P}(G = 2|X = x) = \frac{1}{1 + e^{\beta_0 + \beta_1 x}}$$

which of course has $\mathbb{P}(G = 1|X = x) + \mathbb{P}(G = 2|X = x) = 1$

The decision boundary is the set of points for which logit are zero. This is the hyperplane defined by

$$\{x : \beta_0 + \beta_1 x = 0\}$$

As a reminder, to perform classification, we need the posterior probabilities $\mathbb{P}(G|X)$ for each class

Let $f_k(x)$ be the conditional density of X for class k

Let π_k be the prior probability class of k with $\sum_{k=1}^{K} \pi_k = 1$

Using Bayes' theorem we have:

$$\mathbb{P}(G = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{\ell=1}^{K} f_\ell(x)\pi_\ell}$$

So we need $f_k(x)$

- linear and quadratic discriminant analysis use Gaussian densities
- more flexible mixtures of Gaussian allow for nonlinear decision boundaries
- general nonparametric density estimates for each class density allow the most flexibility
- *Naive Bayes* models are a variant of the previous case, and assume that each of the class densities are products of the marginal densities; that is, they assume that the inputs are conditionally independent in each class

[2]

For multivariate Gaussian distributions we have:

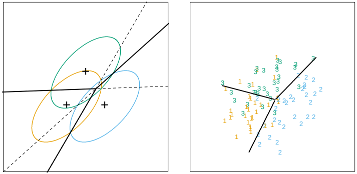
$$f_k(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^p|\boldsymbol{\Sigma_k}|}} e^{\frac{-(\boldsymbol{x}-\boldsymbol{\mu_k})^T\boldsymbol{\Sigma_k}^{-1}(\boldsymbol{x}-\boldsymbol{\mu_k})}{2}}$$

LDA is a special case that arises when all classes have the same covariance matrix ($\boldsymbol{\Sigma_k} = \boldsymbol{\Sigma}, \forall k$)

For this case, the log-ratio (similar to logit) function is:

$$\begin{aligned}
\log \frac{\mathbb{P}(G=k|X=x)}{\mathbb{P}(G=\ell|X=x)} &= \log \frac{f_k(x)}{f_\ell(x)} + \log \frac{\pi_k}{\pi_\ell} \\
&= \log \frac{\pi_k}{\pi_\ell} - \frac{1}{2}(\boldsymbol{\mu_k}+\boldsymbol{\mu_\ell})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu_k}-\boldsymbol{\mu_\ell}) + \boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu_k}-\boldsymbol{\mu_\ell}) \\
&= \boldsymbol{a} + \boldsymbol{x}^T\boldsymbol{b}
\end{aligned}$$

So this is clearly a linear function of $\boldsymbol{x}$



[2]

FIGURE 4.5. *The left panel shows three Gaussian distributions, with the same covariance and different means. Included are the contours of constant density enclosing 95% of the probability in each case. The Bayes decision boundaries between each pair of classes are shown (broken straight lines), and the Bayes decision boundaries separating all three classes are the thicker solid lines (a subset of the former). On the right we see a sample of* 30 *drawn from each Gaussian distribution, and the fitted LDA decision boundaries.*

Linear Discriminant Functions

An equivalent expression to maximizing the posterior probabilities, is to choose the maximum of the *linear discriminant functions* given by:

$$\delta_k(x) = \boldsymbol{x}^T \Sigma^{-1} \boldsymbol{\mu_k} - \frac{1}{2} \boldsymbol{\mu_k}^T \Sigma^{-1} \boldsymbol{\mu_k} + \log \pi_k$$

with $G(x) = \arg\max_k \delta_k(x)$.

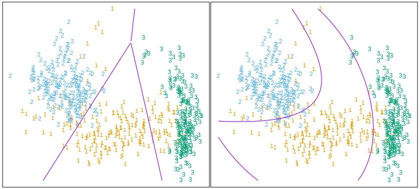
In practice, we don't know the parameters for the multivariate normal distribution. So instead we work with the estimated values, taken from our training data:

- $$\hat{\pi}_k = \frac{N_k}{N}$$
  where N_k is the number of class-k observations
- $$\hat{\mu}_k = \sum_{g_i=k} \frac{x_i}{N_k}$$
- $$\hat{\Sigma} = \sum_{k=1}^{K} \sum_{g_i=k} \frac{(x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T}{N - K}$$

Quadratic Discriminant Analysis (QDA)

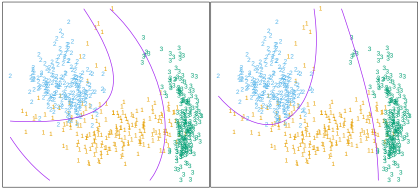
Let's relax the assumption that all covariance matrices are the same. This means that each class has its own covariance matrix Σ_k and so the discriminant function is no longer *linear*. We have our new discriminant function as:

$$\delta_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu_k})^T \Sigma_k^{-1}(\boldsymbol{x} - \boldsymbol{\mu_k}) + \log \pi_k$$



[2]

FIGURE 4.1. *The left plot shows some data from three classes, with linear decision boundaries found by linear discriminant analysis. The right plot shows quadratic decision boundaries. These were obtained by finding linear boundaries in the five-dimensional space* $X_1, X_2, X_1X_2, X_1^2, X_2^2$*. Linear inequalities in this space are quadratic inequalities in the original space.*



[2] **FIGURE 4.6.** *Two methods for fitting quadratic boundaries. The left plot shows the quadratic decision boundaries for the data in Figure 4.1 (obtained using LDA in the five-dimensional space $X_1, X_2, X_1X_2, X_1^2, X_2^2$). The right plot shows the quadratic decision boundaries found by QDA. The differences are small, as is usually the case.*

Imagine that you have a sequence of n iid samples from a known $N(\mu, \sigma^2)$ distribution. The probability density that you would get sample x_i if you knew the mean and variance would be:

$$f(x_i|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$$

Because of the iid nature, the conditional pdf for the entire sample would be:

$$f(x_1, x_2, \ldots, x_n|\mu,\sigma) = \prod_{i=1}^{n} f(x_i|\mu,\sigma)$$

How would this process work in reverse?

We would still have the conditional pdf as:

$$f(x_1, x_2, \ldots, x_n|\mu, \sigma) = \prod_{i=1}^{n} f(x_i|\mu, \sigma)$$

but now we don't know μ and σ so we instead we interpret as a *likelihood function*:

$$\mathcal{L}(\mu, \sigma|x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i)$$

The purpose of the MLE is to find the parameters that maximize the likelihood function. We can express this using vector notation as:

$$\mathcal{L}(\beta|\boldsymbol{x}) = f(\boldsymbol{x}|\beta)$$

where β is the parameter vector and $\boldsymbol{x}$ is the data vector

A slight modification is to find the parameters that maximize the log of the likelihood function given by:

$$l(\beta|\boldsymbol{x}) = \frac{1}{n}\log\mathcal{L}(\beta|\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n}\log f(x_i|\beta)$$

Consider the original problem. How would we determine the parameters for the normal distribution for our *n* observed data points? The log-likelihood function is:

$$l(\hat{\mu}, \hat{\sigma}|\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} \log\left[\frac{1}{\sqrt{2\pi\hat{\sigma}^2}} e^{\frac{-(x_i-\hat{\mu})^2}{2\hat{\sigma}^2}}\right]$$

which simplifies to

$$l(\hat{\mu}, \hat{\sigma}|\boldsymbol{x}) = \log\left(\frac{1}{\sqrt{2\pi\hat{\sigma}^2}}\right) - \frac{1}{2n}\sum_{i=1}^{n}\left(\frac{x_i - \hat{\mu}}{\hat{\sigma}}\right)^2$$

To determine the parameters that maximize this, we look to the partial derivatives:

$$\frac{\partial l}{\partial \hat{\mu}} = -\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \hat{\mu}}{\hat{\sigma}}\right) = 0$$

$$\frac{\partial l}{\partial \hat{\sigma}} = -\frac{1}{\hat{\sigma}} + \frac{1}{n\hat{\sigma}^3}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 = 0$$

These will have solutions:

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

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2}$$

The likelihood function would then be:

$$\mathcal{L} = \prod_{i=1}^{n} p(x_i)$$

but with $p(x) = \mathbb{P}(y = 1|x)$ and $1 - p(x) = \mathbb{P}(y = 0|x)$ we have

$$\mathcal{L} = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} (1 - p(x_i))$$

We maximize this function to determine β_0 and β_1

We generalize this for a scenario where we have K classes when dealing with the logit functions as:

$$\log \frac{\mathbb{P}(G = 1|X = x)}{\mathbb{P}(G = K|X = x)} = \beta_{10} + \beta_1^T x$$
$$\log \frac{\mathbb{P}(G = 2|X = x)}{\mathbb{P}(G = K|X = x)} = \beta_{20} + \beta_2^T x$$
$$\vdots$$
$$\log \frac{\mathbb{P}(G = K-1|X = x)}{\mathbb{P}(G = K|X = x)} = \beta_{(K-1)0} + \beta_{K-1}^T x$$

This leads to the probabilities: For $k = 1, \ldots, K-1$:

$$\mathbb{P}(G = k|X = x) = \frac{e^{\beta_{k0} + \beta_k^T x}}{1 + \sum_{\ell=1}^{K} e^{\beta_{\ell 0} + \beta_\ell^T x}}$$

and for K:

$$\mathbb{P}(G = K|X = x) = \frac{1}{1 + \sum_{\ell=1}^{K} e^{\beta_{\ell 0} + \beta_\ell^T x}}$$

[1] Trevor Hastie Gareth James, Daniela Witten and Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R*. Number v. 6. Springer, 2013.

[2] Robert Tibshirani Trevor Hastie and Jerome Friedman. *The Elements of Stastical Learning: Data Mining, Inference, and Prediction*. Number v.2 in Springer Series in Statistics. Springer, 2009.