Linear and Quadratic Discriminant Analysis

Colan F. Biemer

Linear Discriminant Analysis (LDA) models $P[X = x \,|\, Y = k]$ and $P[Y = k]$. The latter formula is easy to calculate.

$$P[Y = k] = \frac{n_k}{n}$$

Where n is the total number of input rows in the dataset and n_k is the number of rows associated with class k. The former can be used with Bayes Rule.

$$\begin{aligned} P[Y = k \mid X = x] &= \frac{P[X = x \mid Y = k]\, P[Y = k}{P[X = x]} \\ &= \frac{P[X = x \mid Y = k]\, P[Y = k]}{\sum_{j=1}^{K} P[Y = j]\, P[Y = j \mid X = x]} \end{aligned}$$

Where K is the total number of classes. This brings us to the question of: how we are going to model $P[X = x|Y = k]$? LDA assumes that the dataset fits a normal distribution [1] and it calculates for each column the mean, μ_k. It does not, though, calculate a σ_k^2 for each column. Instead, it assumes that $\sigma_1^2 = \sigma_2^2 = ... = \sigma_K^2$. For now assume that we have one input column, $p = 1$.

$$\begin{aligned} \mu_k &= \frac{1}{n_k} \sum_{i:Y_i=k} x_i \\ \sigma^2 &= \frac{1}{n-K} \sum_{k=1}^{k} \sum_{i:Y_i=k} (x_i - \mu_k)^2 \end{aligned}$$

With these two equations, we've calculated what we need to find the probability that a class is k given input x.

$$P[Y = k \mid X = x] = \frac{P[Y = k] \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ \frac{-1}{2} \left(\frac{x - \mu_k}{\sigma} \right)^2 \right\}}{\sum_{j=1}^{K} P[Y = j] \exp\left\{ \frac{-1}{2} \left(\frac{x - \mu_j}{\sigma} \right)^2 \right\}}$$

We can simplify this by taking the log. When we do that, we choose the $Y = k$ with the largest output. But, this can be easily looked up and we won't spend time going through the calculation. Instead, we'll look to the case of $p > 1$. The way we calculated σ^2 is no longer correct, we need to calculate the covariance matrix Σ [2]. Rather than going into what is an easy google search, let's look at the normal distribution for $p > 1$.

$$N(X \mid \mu, \Sigma) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\left\{ \frac{-1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right\}$$

At this point, it should be clear that LDA differs from naive bayes [3] because it does not assume independence. We start with Bayes but go in a very different direction. We can again simplify by taking the log and we won't go through the calculation.

[1] `https://en.wikipedia.org/wiki/Normal_distribution`
[2] `https://datascienceplus.com/understanding-the-covariance-matrix/`
[3] `https://bi3mer.github.io/blog/post_31/biemer_naivebayes.pdf`

That is LDA and now we move onto quadratic discriminant analysis (QDA). The part that bothered me about LDA was that we don't find a variance for each class. QDA does exactly this. It finds a Σ_k, and otherwise there is no difference between the methods. This, though, comes at a cost. LDA finds $p(p+1)/2$ parameters. There are p mean values and the covariance matrix is a $P \times P$ matrix. However, we do not need to know all the values in the matrix. Take the case of $p = 3$.

$$\begin{bmatrix} cov(x,x) & cov(x,y) & cov(x,z) \\ \cancel{cov(y,x)} & cov(y,y) & cov(x,z) \\ \cancel{cov(z,x)} & \cancel{cov(z,y)} & cov(z,z) \end{bmatrix}$$

We can cancel out cases like $cov(y, x)$ since covariance is a symmetric property, $cov(y, x) = cov(x, y)$. We don't have to store the result twice. This is why we have $p(p + 1)/2$ parameters for LDA instead of $P^2 + P$. QDA has more parameters since we calculate a separate covariance matrix for each class: $Kp(p + 1)/2$.

What this tells us is that QDA is going to have more variance and less bias than LDA. We can predict this based on the number of parameters but we can also think about the math. LDA calculates a covariance matrix for all classes whereas QDA calculates one for each class. The bias for LDA is that the covariance is the same across classes where as QDA does not have that bias. We can also use the equations to know that LDA will require less data than QDA. QDA is going to be more likely to overfit than LDA since LDA has a larger bias. Though note that both have a bias of assuming a normal distribution.

This work is based off readings from several text books [1, 2, 3, 4]. The goal is simplify LDA and QDA to be as approachable as is reasonable.

References

[1] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[2] J. Friedman, T. Hastie, R. Tibshirani *et al.*, *The elements of statistical learning*. Springer series in statistics New York, 2001, vol. 1, no. 10.

[3] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[4] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*. Springer, 2013, vol. 112.