Gaussian Discriminant Analyses

compiled by Alvin Wan from Professor Benjamin Recht's lecture

Let us briefly survey all possibilities available; we have three ways to create classifiers.

1. **Empirical Risk Minimization (ERM)**

   We may choose ridge regression or lasso, for example. In general, we would pick an objective function of the following form.

   $$\text{minimize}_w \frac{1}{n} \sum \text{loss}(w; (x_i, y_i)) + \lambda \text{penalty}$$

2. **Generative Models**

   This is a two-step process. For each class c_i, fit the probability of the provided data x.

   $$\Pr(X|Y = c_i)$$

   Using this, estimate $\Pr(Y = c)$ and then make predictions from x. To minimize the probability of error, pick y to maximize the following

   $$\Pr(Y|X)$$

   Note that in the interim, with the first step, we have a viable way to *generate* xs. This could be a goal in and of itself: our mission could be to generate scholarly articles or computer programs, for example. On the other hand, this first step may be a particularly difficult problem to solve.

3. **Discriminative Models**

   We directly model the class conditional probabilities, directly.

   $$\Pr(Y|X)$$

In this note, we will focus on two types of generative models, both variants of Gaussian Discriminant Analysis.

1 Gaussian Discriminant Analysis

We will assume that each conditional is Gaussian and that the set $C = \{i|y_i = c\}$ is comprised of the indices for all y_i of this class.

$$\Pr(X|Y = c) \sim \mathcal{N}(\mu_c, \sigma^2 I)$$

We will model $\Pr(X|Y = c)$ using maximum likelihood estimates of the Guassian.

$$\hat{\mu}_c = \frac{1}{|C|}\sum_{i \in C} x_i$$

$$\hat{\Lambda}_c = \frac{1}{|C|}\sum_{i \in C}(x_i - \hat{\mu}_c)(x_i - \hat{\mu}_c)^T$$

Using this, we find that the likelihood of any class is directly proportional to the number of items in that class. In other words,

$$\Pr(Y = c) = \frac{|C|}{n}$$

1.1 Decision Rule

We choose the class that maximizes the joint probability $\Pr(X, Y) = \Pr(X|Y)\Pr(Y)$.

$$\text{argmax}\Pr(X|Y = c)\Pr(Y = c)$$

Plugging in for both the conditional and the density, we have the following.

$$\text{argmax}(2\pi)^{k/2}|\hat{\Lambda}_c|^{1/2}\text{exp}(-\frac{1}{2}(x_i - \hat{\mu}_c)^T\hat{\Lambda}_c^{-1}(x_i - \hat{\mu}_c))\frac{|C|}{n}$$

Since the log function is monotonically increasing, we can equivalently pick according to the maximum of these values logged.

$$\underset{c}{\operatorname{argmax}} -\frac{1}{2}(x-\hat{\mu}_c)^T\hat{\Lambda}_c^{-1}(x-\hat{\mu}_c) - \frac{1}{2}\log(2\pi)^k|\hat{\Lambda}_c| + \log\frac{|C|}{n}$$

Finally, negate the optimization function so that the maximization problem now becomes a minimization problem.

$$\underset{c}{\operatorname{argmin}} \frac{1}{2}(x-\hat{\mu}_c)^T\hat{\Lambda}_c^{-1}(x-\hat{\mu}_c) + \frac{1}{2}\log(2\pi)^k|\hat{\Lambda}_c| - \log\frac{|C|}{n}$$

To find our class, we simply evaluate the quadratics and pick the c that corresponds to our smallest value.

2 Quadratic Discriminant Analysis (QDA)

Quadratic Discriminant Analysis is a more general version of a linear classifier. The quadratic term allows QDA to separate data using a quadric surface in higher dimensions. For the two-class case, the decision boundary lies along all x such that $Q_1(x) = Q_2(x)$, where each Q_c is the following.

$$Q_c(x) = -\frac{1}{2}(x-\hat{\mu}_c)^T\hat{\Lambda}_c^{-1}(x-\hat{\mu}_c) - \frac{1}{2}\log(2\pi)^k|\hat{\Lambda}_c| + \log\frac{|C|}{n}$$

This generative model yields $\Pr(Y|X)$, allowing us to quantify the confidence of our classification. To simplify the expression, let C be the event that $Y = c$ and $\bar{C}$ be the event that $Y \neq c$.

$$\begin{aligned}\Pr(C|X) &= \frac{\Pr(X|C)\Pr(C)}{\Pr(X|C)\Pr(C) + \Pr(X|\bar{C})\Pr(\bar{C})} \\ &= \frac{1}{1 + \frac{\Pr(X|\bar{C})\Pr(\bar{C})}{\Pr(X|C)\Pr(C)}}\end{aligned}$$

We consider the fraction in the denominator $\frac{\Pr(X|\bar{C})\Pr(\bar{C})}{\Pr(X|C)\Pr(C)}$. We will use two simplifications. First, $\alpha = e^{\log \alpha}$, and second, $\frac{e^\alpha}{e^\beta} = e^{\alpha-\beta}$. This leads allows us to express the fraction as a function of $Q_i(x)$.

$$\Pr(C|X) = \frac{1}{1 + \exp(Q_{\bar{C}} - Q_C)}$$

For the two class case, we have that on the boundary $Q_C = Q_{\bar{C}}$ for either class. Thus,

$$\Pr(Y = c_1|X) = \Pr(Y = c_2|X) = \frac{1}{1 + e^0} = \frac{1}{2}$$

3 Linear Discriminant Analysis (LDA)

In LDA, we assume that covariance matrices Λ_i are the same for all classes.

$$\hat{\Lambda} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j \in C_k} (x_j - \hat{\mu}_i)(x_j - \hat{\mu}_i)^T$$

Let us consider the general class-conditional expression derived in the analysis of QDA. In our two class case, for our two classes C_1 and C_2, we observe that the exponent in the denominator $Q_1 - Q_2$ is the following.

$$\begin{aligned} &= -\frac{1}{2}(x - \hat{\mu}_1)^T \hat{\Lambda}^{-1}(x - \hat{\mu}_1) - \frac{1}{2}\log(2\pi)^k|\hat{\Lambda}| + \log \frac{|C_1|}{n} \\ &\quad + \frac{1}{2}(x - \hat{\mu}_2)^T \hat{\Lambda}^{-1}(x - \hat{\mu}_2) + \frac{1}{2}\log(2\pi)^k|\hat{\Lambda}| - \log \frac{|C_2|}{n} \end{aligned}$$

We can first re-arrange terms, then combine the constants that are not a function of x.

$$
\begin{aligned}
=& -\frac{1}{2}(x-\hat{\mu}_1)^T\hat{\Lambda}^{-1}(x-\hat{\mu}_1) + \frac{1}{2}(x-\hat{\mu}_2)^T\hat{\Lambda}^{-1}(x-\hat{\mu}_2) \\
& -\frac{1}{2}\log(2\pi)^k|\hat{\Lambda}| + \log\frac{|C_1|}{n} + \frac{1}{2}\log(2\pi)^k|\hat{\Lambda}| - \log\frac{|C_2|}{n} \\
=& -\frac{1}{2}(x-\hat{\mu}_1)^T\hat{\Lambda}^{-1}(x-\hat{\mu}_1) + \frac{1}{2}(x-\hat{\mu}_2)^T\hat{\Lambda}^{-1}(x-\hat{\mu}_2) + \log\frac{|C_1|}{|C_2|}
\end{aligned}
$$

Now, expand the first two terms.

$$
\begin{aligned}
=& -\frac{1}{2}x^T\hat{\Lambda}^{-1}x + \hat{\mu}_1^T\hat{\Lambda}^{-1}x - \frac{1}{2}\hat{\mu}_1^T\hat{\Lambda}^{-1}\hat{\mu}_1 + \frac{1}{2}x^T\hat{\Lambda}^{-1}x - \hat{\mu}_2^T\hat{\Lambda}^{-1}x + \frac{1}{2}\hat{\mu}_2^T\hat{\Lambda}^{-1}\hat{\mu}_2 \\
& + \log\frac{|C_1|}{|C_2|}
\end{aligned}
$$

We can cancel out the two $\frac{1}{2}x^T\hat{\Lambda}^{-1}x$.

$$
=\hat{\mu}_1^T\hat{\Lambda}^{-1}x - \frac{1}{2}\hat{\mu}_1^T\hat{\Lambda}^{-1}\hat{\mu}_1 - \hat{\mu}_2^T\hat{\Lambda}^{-1}x + \frac{1}{2}\hat{\mu}_2^T\hat{\Lambda}^{-1}\hat{\mu}_2 + \log\frac{|C_1|}{|C_2|}
$$

We finally re-arrange to get our desired expression.

$$
=(\hat{\mu}_1-\hat{\mu}_2)^T\hat{\Lambda}^{-1}x - \frac{1}{2}\hat{\mu}_1^T\hat{\Lambda}^{-1}\hat{\mu}_1 + \frac{1}{2}\hat{\mu}_2^T\hat{\Lambda}^{-1}\hat{\mu}_2 + \log\frac{|C_1|}{|C_2|}
$$

We find that this is equivalent to $\beta^T x + \alpha$, where

$$
\begin{aligned}
\beta &= \hat{\mu}_1 - \hat{\mu}_2 \\
\alpha &= -\frac{1}{2}\hat{\mu}_1^T\hat{\Lambda}^{-1}\hat{\mu}_1 + \frac{1}{2}\hat{\mu}_2^T\hat{\Lambda}^{-1}\hat{\mu}_2 + \log\frac{|C_1|}{|C_2|}
\end{aligned}
$$

As a result, this decision boundary is linear.

4 Special Cases

In one case, we have a spherical $\Lambda = \sigma^2 I$. We see the following decision boundary. First, take our expression from the last section and plug in Λ.

$$(\hat{\mu}_1 - \hat{\mu}_2)^T \sigma^2 I^{-1} x - \frac{1}{2}\hat{\mu}_1^T \sigma^2 I^{-1} \hat{\mu}_1 + \frac{1}{2}\hat{\mu}_2^T \sigma^2 I^{-1} \hat{\mu}_2 = 0$$

Multiply all terms by σ^2.

$$(\hat{\mu}_1 - \hat{\mu}_2)^T x - \frac{1}{2}\hat{\mu}_1^T \hat{\mu}_1 + \frac{1}{2}\hat{\mu}_2^T \hat{\mu}_2 = 0$$
$$(\hat{\mu}_1 - \hat{\mu}_2)^T x - \frac{1}{2}(\hat{\mu}_1^T \hat{\mu}_1 - \hat{\mu}_2^T \hat{\mu}_2) = 0$$

Finally, note that $a^2 - b^2 = (a+b)(a-b)$, and combine like terms.

$$(\hat{\mu}_1 - \hat{\mu}_2)^T x - \frac{1}{2}(\hat{\mu}_1 - \hat{\mu}_2)^T(\hat{\mu}_1 + \hat{\mu}_2) = 0$$
$$(\hat{\mu}_1 - \hat{\mu}_2)^T (x - \frac{1}{2}(\hat{\mu}_1 + \hat{\mu}_2)) = 0$$

In the second case, we do not have a spherical Λ. Thus, we see the following decision boundary.

$$(\hat{\mu}_0 - \hat{\mu}_1)^T \Lambda^{-1} (x - \frac{1}{2}(\hat{\mu}_0 + \hat{\mu}_1)) = 0$$

Both of these formulations have an intuitive interpretation. We are effectively taking the midpoint of the two means but in a vector space.