CS 189 Introduction to Machine Learning
Spring 2018 Note 18

1 Gaussian Discriminant Analysis

Recall the idea of **generative models**: we classify an arbitrary datapoint $\mathbf{x}$ with the class label that maximizes the joint probability $p(\mathbf{x}, Y)$ over the label Y:

$$\hat{y} = \arg\max_k \ p(\mathbf{x}, Y = k)$$

Generative models typically form the joint distribution by explicitly forming the following:

- A prior probability distribution over all classes:

$$P(k) = P(\text{class} = k)$$

- A conditional probability distribution for each class $k \in \{1, 2, ..., K\}$:

$$p_k(\mathbf{X}) = p(\mathbf{X}|\text{class } k)$$

In total there are $K + 1$ probability distributions: 1 for the prior, and K for all of the individual classes. Note that the prior probability distribution is a categorical distribution over the K discrete classes, whereas each class conditional probability distribution is a continuous distribution over $\mathbb{R}^d$ (often represented as a Gaussian). Using the prior and the conditional distributions in conjunction, we have (from Bayes' rule) that maximizing the joint probability over the class labels is equivalent to maximizing the posterior probability of the class label:

$$\hat{y} = \arg\max_k \ p(\mathbf{x}, Y = k) = \arg\max_k \ P(k)\, p_k(\mathbf{x}) = \arg\max_k \ P(Y = k|\mathbf{x})$$

Consider the example of digit classification. Suppose we are given dataset of images of handwritten digits each with known values in the range $\{0, 1, 2, \ldots, 9\}$. The task is, given an image of a handwritten digit, to classify it to the correct digit. A generative classifier for this this task would effectively form a the prior distribution and conditional probability distributions over the 10 possible digits and choose the digit that maximizes posterior probability:

$$\hat{y} = \arg\max_{k \in \{0,1,2...,9\}} \ p(\text{digit} = k|\text{image}) = \arg\max_{k \in \{0,1,2...,9\}} \ P(\text{digit} = k)\, p(\text{image}|\text{digit} = k)$$

Maximizing the posterior will induce regions in the feature space in which one class has the highest posterior probability, and **decision boundaries** in between classes where the posterior probability of two classes are equal.

Gaussian Discriminant Analysis (GDA) is a specific generative method in which the class conditional probability distributions are Gaussian: $(\mathbf{X}|Y = k) \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. (Caution: the term "discriminant" in GDA is misleading; GDA is a generative method, it is not a discriminative method!)

Assume that we are given a training set $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ of n points. Estimating the prior distribution is the same for any other generative model. The probability of a class k is

$$P(k) = \frac{n_k}{n}$$

where n_k is the number of training points that belong to class k. We can estimate the parameters of the conditional distributions with MLE. Once we form the estimated prior conditional distributions, we use Bayes' Rule to directly solve the optimization problem

$$\begin{aligned}
\hat{y} &= \arg\max_k \; p(k \mid \mathbf{x}) \\
&= \arg\max_k \; P(k)\; p_k(\mathbf{x}) \\
&= \arg\max_k \; \ln\Big(P(k)\Big) + \ln\Big((\sqrt{2\pi})^d p_k(\mathbf{x})\Big) \\
&= \arg\max_k \; \ln\Big(P(k)\Big) - \frac{1}{2}(\mathbf{x} - \hat{\boldsymbol{\mu}}_k)^\top \hat{\boldsymbol{\Sigma}}_k^{-1}(\mathbf{x} - \hat{\boldsymbol{\mu}}_k) - \frac{1}{2}\ln\Big(|\hat{\boldsymbol{\Sigma}}_k|\Big) = Q_k(\mathbf{x})
\end{aligned}$$

For future reference, let's use $Q_k(\mathbf{x}) = \ln\Big(\sqrt{2\pi}\Big)^d P(k)\; p_k(\mathbf{x})$ to simplify our notation.

We classify an arbitrary test point

$$\hat{y} = \arg\max_{k \in \{1,2,\ldots,K\}} \; Q_k(\mathbf{x})$$

GDA comes in two flavors: Quadratic Discriminant Analysis (QDA) in which the decision boundary is quadratic, and Linear Discriminant Analysis (LDA) in which the decision boundary is linear. We will now present both and compare them in detail.

1.1 QDA Classification

In **Quadratic Discriminant Analysis (QDA)**, the class conditional probability distributions are independent Gaussians — namely, the covariance $\boldsymbol{\Sigma}_k$ of class k has no dependence/relation to that of the other classes.

Due to this independence property, we can estimate the true mean and covariance $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ for each class conditional probability distribution $p_k(\mathbf{X})$ independently, with the n_k samples in our training data that are classified as class k. The MLE estimate for the parameters of $p_k(\mathbf{X})$ is:

$$\begin{aligned}
\hat{\boldsymbol{\mu}}_k &= \frac{1}{n_k} \sum_{i:y_i=k} \mathbf{x}_i \\
\hat{\boldsymbol{\Sigma}}_k &= \frac{1}{n_k} \sum_{i:y_i=k} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)^\top
\end{aligned}$$

1.2 LDA Classification

While QDA is a reasonable approach to classification, we might be interested in simplifying our model to reduce the number of parameters we have to learn. One way to do this is through **Linear Discriminant Analysis (LDA)** classification. Just as in QDA, LDA assumes that the class conditional probability distributions are normally distributed with different means $\boldsymbol{\mu}_k$, but LDA is different from QDA in that it requires all of the distributions to share the same covariance matrix $\boldsymbol{\Sigma}$. This is a simplification which, in the context of the Bias-Variance tradeoff, increases the bias of our method but may help decrease the variance.

The training and classification procedures for LDA are almost identical that of QDA. To compute the within-class means, we still want to take the empirical mean. However, the empirical covariance for all classes is now computed as

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{y_i})(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{y_i})^\top$$

One way to understand this formula is as a weighted average of the within-class covariances. Here, assume we have sorted our training data by class and we can index through the $\mathbf{x}_i$'s by specifying a class k and the index within that class j:

$$\begin{aligned}
\hat{\boldsymbol{\Sigma}} &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{y_i})(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{y_i})^\top \\
&= \frac{1}{n}\sum_{k=1}^{K}\sum_{i:y_i=k}^{n_k}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)^\top \\
&= \frac{1}{n}\sum_{k=1}^{K} n_k\, \boldsymbol{\Sigma}_k \\
&= \sum_{k=1}^{K}\frac{n_k}{n}\boldsymbol{\Sigma}_k
\end{aligned}$$

1.3 LDA and QDA Decision Boundary

Let's now derive the form of the decision boundary for QDA and LDA. As we will see, the term *quadratic* in QDA and *linear* in LDA actually signify the shape of the decision boundary. We will prove this claim using binary (2-class) examples for simplicity (class A and class B). An arbitrary point $\mathbf{x}$ is classified according to the following cases:

$$\hat{y} = \begin{cases} A & Q_A(\mathbf{x}) > Q_B(\mathbf{x}) \\ B & Q_A(\mathbf{x}) < Q_B(\mathbf{x}) \\ \text{Either } A \text{ or } B & Q_A(\mathbf{x}) = Q_B(\mathbf{x}) \end{cases}$$

The decision boundary is the set of all points in $\mathbf{x}$-space that are classified according to the third case.

1.3.1 Identical Conditional Distributions with Identical Priors

The simplest case is when the two classes are equally likely in prior, and their conditional probability distributions are isotropic with identical covariances. Recall that isotropic Gaussian distributions have covariances of the form of $\mathbf{\Sigma} = \sigma^2\mathbf{I}$, which means that their isocontours are circles. In this case, $p_A(\mathbf{X})$ and $p_B(\mathbf{X})$ have identical covariances of the form $\mathbf{\Sigma}_A = \mathbf{\Sigma}_B = \sigma^2\mathbf{I}$.

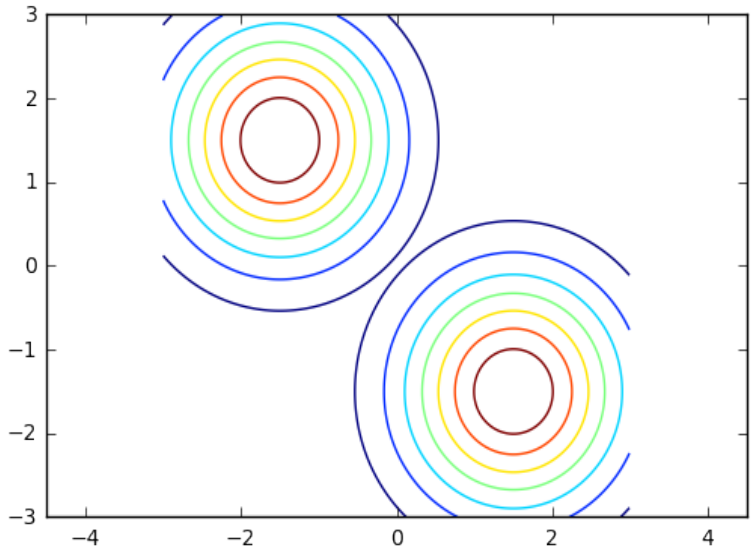

Figure 1: Contour plot of two isotropic, identically distributed Gaussians in $\mathbb{R}^2$. The circles are the level sets of the Gaussians.

Geometrically, we can see that the task of classifying a 2-D point into one of the two classes amounts simply to figuring out which of the means it's closer to. Using our notation of $Q_k(\mathbf{x})$ from before, this can be expressed mathematically as:

$$
\begin{aligned}
Q_A(\mathbf{x}) &= Q_B(\mathbf{x}) \\
\ln\big(P(A)\big) - \frac{1}{2}(\mathbf{x}-\hat{\boldsymbol{\mu}}_A)^\top\hat{\mathbf{\Sigma}}_A^{-1}(\mathbf{x}-\hat{\boldsymbol{\mu}}_A) - \frac{1}{2}\ln\Big(|\hat{\mathbf{\Sigma}}_A|\Big) &= \ln\big(P(B)\big) - \frac{1}{2}(\mathbf{x}-\hat{\boldsymbol{\mu}}_B)^\top\hat{\mathbf{\Sigma}}_B^{-1}(\mathbf{x}-\hat{\boldsymbol{\mu}}_B) - \frac{1}{2}\ln\Big(|\hat{\mathbf{\Sigma}}_B|\Big) \\
\ln\left(\frac{1}{2}\right) - \frac{1}{2}(\mathbf{x}-\hat{\boldsymbol{\mu}}_A)^\top\sigma^{-2}\mathbf{I}(\mathbf{x}-\hat{\boldsymbol{\mu}}_A) - \frac{1}{2}\ln\Big(|\sigma^2\mathbf{I}|\Big) &= \ln\left(\frac{1}{2}\right) - \frac{1}{2}(\mathbf{x}-\hat{\boldsymbol{\mu}}_B)^\top\sigma^{-2}\mathbf{I}(\mathbf{x}-\hat{\boldsymbol{\mu}}_B) - \frac{1}{2}\ln\Big(|\sigma^2\mathbf{I}|\Big) \\
(\mathbf{x}-\hat{\boldsymbol{\mu}}_A)^\top(\mathbf{x}-\hat{\boldsymbol{\mu}}_A) &= (\mathbf{x}-\hat{\boldsymbol{\mu}}_B)^\top(\mathbf{x}-\hat{\boldsymbol{\mu}}_B)
\end{aligned}
$$

The decision boundary is the set of points $\mathbf{x}$ for which $\|\mathbf{x}-\hat{\boldsymbol{\mu}}_A\|_2 = \|\mathbf{x}-\hat{\boldsymbol{\mu}}_B\|_2$, which is simply the set of points that are equidistant from $\hat{\boldsymbol{\mu}}_A$ and $\hat{\boldsymbol{\mu}}_B$. This decision boundary is linear because the set of points that are equidistant from $\hat{\boldsymbol{\mu}}_A$ and $\hat{\boldsymbol{\mu}}_B$ are simply the perpendicular bisector of the segment connecting $\hat{\boldsymbol{\mu}}_A$ and $\hat{\boldsymbol{\mu}}_B$.

The next case is when the two classes are equally likely in prior, and their conditional probability distributions are anisotropic with identical covariances.

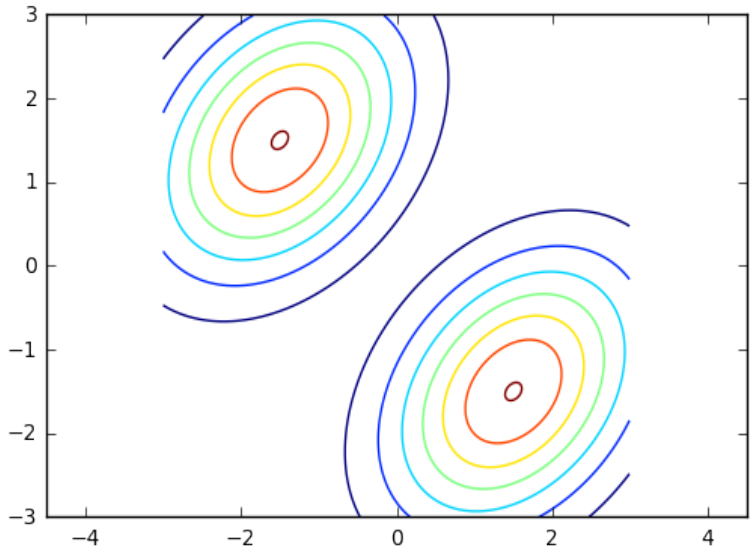

Figure 2: Two anisotropic, identically distributed Gaussians in $\mathbb{R}^2$. The ellipses are the level sets of the Gaussians.

The anisotropic case can be reduced to the isotropic case simply by performing a linear change of coordinates that transforms the ellipses back into circles, which induces a linear decision boundary both in the transformed and original space. Therefore, the decision boundary is still the set of points that are equidistant from $\hat{\boldsymbol{\mu}}_A$ and $\hat{\boldsymbol{\mu}}_B$.

1.3.2 Identical Conditional Distributions with Different Priors

Now, let's find the decision boundary when the two classes still have identical covariances but are not necessarily equally likely in prior:

$$
\begin{aligned}
Q_A(\mathbf{x}) &= Q_B(\mathbf{x}) \\
\ln\big(P(A)\big) - \frac{1}{2}(\mathbf{x}-\hat{\boldsymbol{\mu}}_A)^\top \hat{\boldsymbol{\Sigma}}_A^{-1}(\mathbf{x}-\hat{\boldsymbol{\mu}}_A) - \frac{1}{2}\ln\Big(|\hat{\boldsymbol{\Sigma}}_A|\Big) &= \ln\big(P(B)\big) - \frac{1}{2}(\mathbf{x}-\hat{\boldsymbol{\mu}}_B)^\top \hat{\boldsymbol{\Sigma}}_B^{-1}(\mathbf{x}-\hat{\boldsymbol{\mu}}_B) - \frac{1}{2}\ln\Big(|\hat{\boldsymbol{\Sigma}}_B|\Big) \\
\ln\big(P(A)\big) - \frac{1}{2}(\mathbf{x}-\hat{\boldsymbol{\mu}}_A)^\top \hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}-\hat{\boldsymbol{\mu}}_A) - \frac{1}{2}\ln\Big(|\hat{\boldsymbol{\Sigma}}|\Big) &= \ln\big(P(B)\big) - \frac{1}{2}(\mathbf{x}-\hat{\boldsymbol{\mu}}_B)^\top \hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}-\hat{\boldsymbol{\mu}}_B) - \frac{1}{2}\ln\Big(|\hat{\boldsymbol{\Sigma}}|\Big) \\
\ln\big(P(A)\big) - \frac{1}{2}(\mathbf{x}-\hat{\boldsymbol{\mu}}_A)^\top \hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}-\hat{\boldsymbol{\mu}}_A) &= \ln\big(P(B)\big) - \frac{1}{2}(\mathbf{x}-\hat{\boldsymbol{\mu}}_B)^\top \hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}-\hat{\boldsymbol{\mu}}_B) \\
2\ln\big(P(A)\big) - \mathbf{x}^\top \hat{\boldsymbol{\Sigma}}^{-1}\mathbf{x} + 2\mathbf{x}^\top \hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}_A - \hat{\boldsymbol{\mu}}_A{}^\top \hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}_A &= 2\ln\big(P(B)\big) - \mathbf{x}^\top \hat{\boldsymbol{\Sigma}}^{-1}\mathbf{x} + 2\mathbf{x}^\top \hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}_B - \hat{\boldsymbol{\mu}}_B{}^\top \hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}_B \\
2\ln\big(P(A)\big) + 2\mathbf{x}^\top \hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}_A - \hat{\boldsymbol{\mu}}_A{}^\top \hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}_A &= 2\ln\big(P(B)\big) + 2\mathbf{x}^\top \hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}_B - \hat{\boldsymbol{\mu}}_B{}^\top \hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}_B
\end{aligned}
$$

Simplifying, we have that

$$
\mathbf{x}^\top \hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_A - \hat{\boldsymbol{\mu}}_B) + \left(\ln\left(\frac{P(A)}{P(B)}\right) - \frac{\hat{\boldsymbol{\mu}}_A{}^\top \hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}_A - \hat{\boldsymbol{\mu}}_B{}^\top \hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}_B}{2}\right) = 0
$$

The decision boundary is the level set of a linear function $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} - b$. In fact, the decision boundary is the level set of a linear function (which itself is linear) as long as the two class con-

ditional probability distributions share the same covariance matrices. This is the reason for why LDA has a linear decision boundary.

1.3.3 Nonidentical Conditional Distributions with Different Priors

This is the most general case. We have that:

$$Q_A(\mathbf{x}) = Q_B(\mathbf{x})$$
$$\ln\big(P(A)\big) - \frac{1}{2}(\mathbf{x}-\hat{\boldsymbol{\mu}}_A)^\top \hat{\boldsymbol{\Sigma}}_A^{-1}(\mathbf{x}-\hat{\boldsymbol{\mu}}_A) - \frac{1}{2}\ln\Big(|\hat{\boldsymbol{\Sigma}}_A|\Big) = \ln\big(P(B)\big) - \frac{1}{2}(\mathbf{x}-\hat{\boldsymbol{\mu}}_B)^\top \hat{\boldsymbol{\Sigma}}_B^{-1}(\mathbf{x}-\hat{\boldsymbol{\mu}}_B) - \frac{1}{2}\ln\Big(|\hat{\boldsymbol{\Sigma}}_B|\Big)$$

Here, unlike in LDA when $\boldsymbol{\Sigma}_A = \boldsymbol{\Sigma}_B$, we *cannot* cancel out the quadratic terms in $\mathbf{x}$ from both sides of the equation, and thus our decision boundary is now represented by the level set of an arbitrary quadratic function.

It should now make sense why QDA is short for *quadratic* discriminant analysis and LDA is short for *linear* discriminant analysis!

1.4 LDA and Logistic Regression

As it turns out, LDA and logistic regression share the same type of posterior distribution. We already showed that the posterior distribution in logistic regression is

$$P(Y = A|\mathbf{x}) = \frac{1}{1+e^{\mathbf{w}^\top\mathbf{x}-b}} = s(\mathbf{w}^\top\mathbf{x}-b)$$

for some appropriate vector $\mathbf{w}$ and bias b. Now let's derive the posterior distribution for LDA. From Bayes' rule we have that

$$\begin{aligned} P(Y = A|\mathbf{x}) &= \frac{p(\mathbf{x}|Y=A)P(Y=A)}{p(\mathbf{x}|Y=B)P(Y=B) + p(\mathbf{x}|Y=B)P(Y=B)} \\ &= \frac{e^{Q_A(\mathbf{x})}}{e^{Q_A(\mathbf{x})} + e^{Q_B(\mathbf{x})}} \\ &= \frac{1}{1+e^{Q_A(\mathbf{x})-Q_B(\mathbf{x})}} \end{aligned}$$

We already showed the the decision boundary in LDA is linear — it is the set of points $\mathbf{x}$ such that

$$Q_A(\mathbf{x}) - Q_B(\mathbf{x}) = \mathbf{w}^\top\mathbf{x} - b = 0$$

for some appropriate vector $\mathbf{w}$ and bias b. We therefore have that

$$P(Y = A|\mathbf{x}) = \frac{1}{1+e^{\mathbf{w}^\top\mathbf{x}-b}} = s(\mathbf{w}^\top\mathbf{x}-b)$$

As we can see, even though logistic regression is a discriminative method and LDA is a generative method, both methods complement each other, arriving at the same form for the posterior distribution.

1.5 Generalizing to Multiple Classes

The analysis on the decision boundary in QDA and LDA can be extended to the general case when there are more than two classes. In the multiclass setting, the decision boundary is a collection of linear boundaries in LDA and quadratic boundaries in QDA. The following **Voronoi** diagrams illustrate the point:

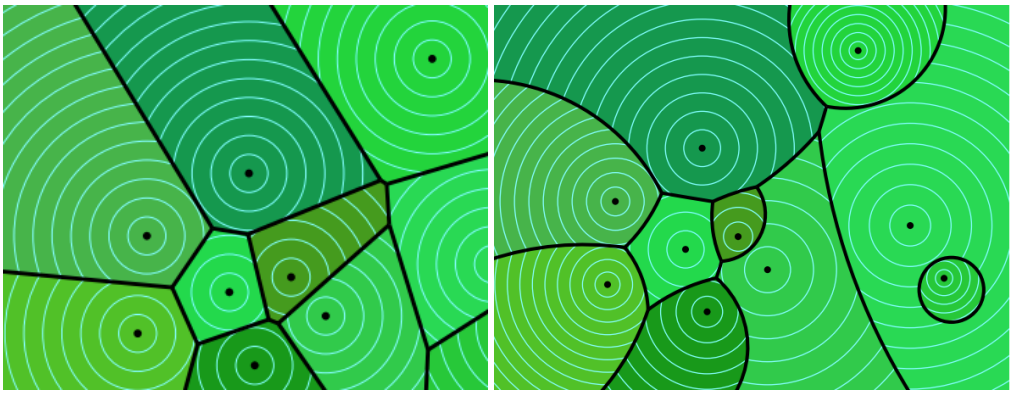

Figure 3: LDA (left) vs QDA (right): a collection of linear vs quadratic level set boundaries. Source: Professor Shewchuk's notes