1 QDA and LDA

1.1 QDA Classification

Quadratic Discriminant Analysis (QDA) is a specific generative method in which the class conditional probability distributions are independent Gaussians: $f_k(.) \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.

(Note: the term "discriminant" in QDA is misleading: remember that QDA is not a discriminative method, it is a generative method!)

For a particular class conditional probability distribution $f_k(.)$, if we do not have the true means and covariances $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$, then our best bet is to estimate them empirically with the samples in our training data that are classified as class k. Recall that the MLE estimate for the parameters of $f_k(.)$ is:

$$\hat{\boldsymbol{\mu}}_k = \frac{1}{n_k} \sum_{t_i=k} \mathbf{x}_i$$
$$\hat{\boldsymbol{\Sigma}}_k = \frac{1}{n_k} \sum_{t_i=k} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)^T$$

Assuming that we know the means and covariances for all the classes, we can use Bayes' Rule to directly solve the optimization problem

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

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

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 is now computed with

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

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:

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

1.3 LDA vs. QDA

Up to this point, we have used the term **quadratic** in QDA and **linear** in LDA. These terms signify the shape of the **decision boundary** in $\mathbf{x}$-space. Given any two classes, the decision boundary represents the points in $\mathbf{x}$-space at which the two classes are equally likely.

Let's study binary (2-class) examples for simplicity. Assume that the two classes in question are class A and class B. An arbitrary point $\mathbf{x}$ can be classified according to three cases:

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

The decision boundary is the set of all points in $\mathbf{x}$-space that are classified according to the third case. Let's look at the form of the decision boundary according to the different scenarios possible under QDA and LDA.

1.3.1 Identical Distributions

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, $f_A(.)$ and $f_B(.)$ have identical covariances of the form $\mathbf{\Sigma}_A = \mathbf{\Sigma}_B = \sigma^2\mathbf{I}$.

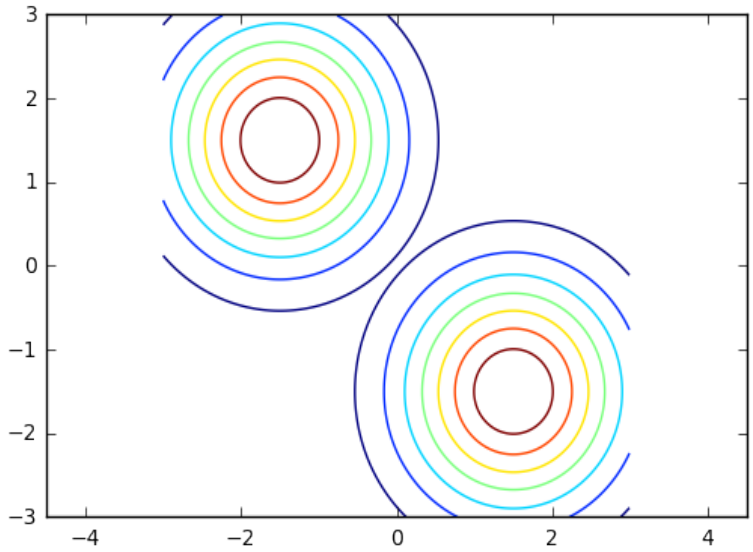

Figure 1: Contour plot of two isotropic, identically distributed Gaussians in $\mathbb{R}^2$. The circles are the level sets of the Gaussians.

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:

$$Q_A(\mathbf{x}) = Q_B(\mathbf{x})$$

$$\ln\left(\frac{1}{2}\right) - \frac{1}{2}(\mathbf{x} - \hat{\boldsymbol{\mu}_A})^T\sigma^2\mathbf{I}(\mathbf{x} - \hat{\boldsymbol{\mu}_A}) - \frac{1}{2}\ln\left(|\sigma^2\mathbf{I}|\right) = \ln\left(\frac{1}{2}\right) - \frac{1}{2}(\mathbf{x} - \hat{\boldsymbol{\mu}_B})^T\sigma^2\mathbf{I}(\mathbf{x} - \hat{\boldsymbol{\mu}_B}) - \frac{1}{2}\ln\left(|\sigma^2\mathbf{I}|\right)$$

$$(\mathbf{x} - \hat{\boldsymbol{\mu}_A})^T(\mathbf{x} - \hat{\boldsymbol{\mu}_A}) = (\mathbf{x} - \hat{\boldsymbol{\mu}_B})^T(\mathbf{x} - \hat{\boldsymbol{\mu}_B})$$

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 $\boldsymbol{\mu}_A$ and $\boldsymbol{\mu}_B$. This decision boundary is linear because the set of points that are equidistant from $\boldsymbol{\mu}_A$ and $\boldsymbol{\mu}_B$ are simply the perpendicular bisector of the segment connecting $\boldsymbol{\mu}_A$ and $\boldsymbol{\mu}_B$.

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

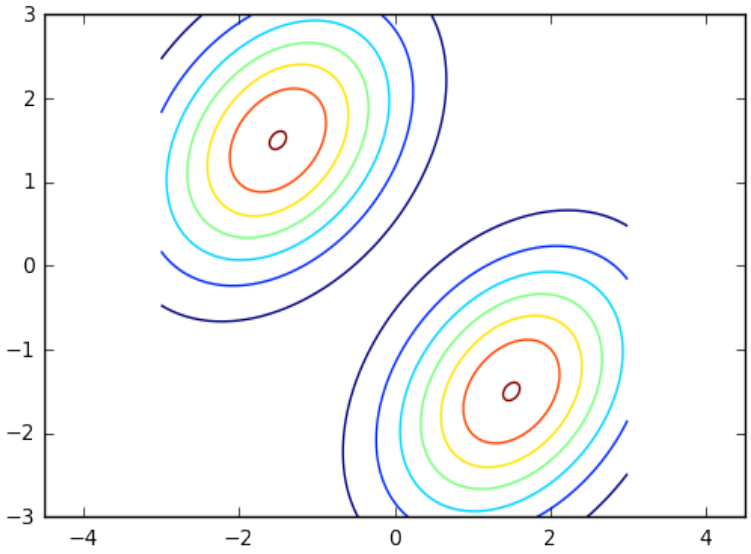

Figure 2: Two anisotropic, identically distributed Gaussians in $\mathbb{R}^2$. The ellipses are the level sets of the Gaussians.

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.

1.3.2 Identical Distributions with Priors

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

$$Q_A(\mathbf{x}) = Q_B(\mathbf{x})$$

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

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

$$2\ln\big(P(A)\big) - \mathbf{x}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{x} + 2\mathbf{x}^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}_A} - \hat{\boldsymbol{\mu}_A}^T\hat{\boldsymbol{\mu}_A} = 2\ln\big(P(B)\big) - \mathbf{x}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{x} + 2\mathbf{x}^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}_B} - \hat{\boldsymbol{\mu}_B}^T\hat{\boldsymbol{\mu}_B}$$

$$2\ln\big(P(A)\big) + 2\mathbf{x}^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}_A} - \hat{\boldsymbol{\mu}_A}^T\hat{\boldsymbol{\mu}_A} = 2\ln\big(P(B)\big) + 2\mathbf{x}^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}_B} - \hat{\boldsymbol{\mu}_B}^T\hat{\boldsymbol{\mu}_B}$$

Simplifying, we have that

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

The decision boundary is the level set of a linear function $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + k$. Notice the pattern: the decision boundary is always the level set of a linear function (which itself is linear) as long as

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

1.3.3 Nonidentical Distributions with Priors

We have that:

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

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!

1.3.4 Generalizing to Multiple Classes

The quadratic nature of the decision boundary in QDA and the linear nature of the decision boundary in LDA still apply to the general case when there are more than two classes. The following excellent figures from Professor Shewchuk's notes illustrate this point:

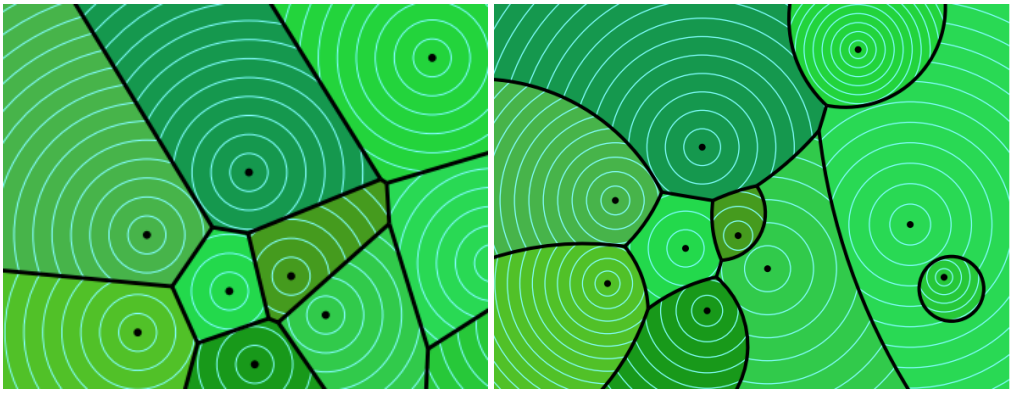

Figure 3: LDA (left) vs QDA (right): a collection of linear vs quadratic level set boundaries