CS 189 Introduction to Machine Learning
Fall 2017 Note 14

The task of **classification** differs from **regression** in that we are now interested in assigning a d-dimensional data point one of a *discrete* number of **classes** instead of assigning it a *continuous* value. Thus, the task is simpler in that there are fewer choices of labels per data point but more complicated in that we now need to somehow factor in information about each class to obtain the classifier that we want.

Here's a formal definition: Given a training set $\mathscr{D} = \{(x_i, t_i)\}_{i=1}^n$ of n points, with each data point $x_i \in \mathbb{R}^d$ paired with a known discrete class label $t_i \in \{1, 2, ..., K\}$, train a classifier which, when fed any arbitrary d-dimensional data point, classifies that data point as one of the K discrete classes.

Classifiers have strong roots in probabilistic modeling. The idea is that given an arbitrary datapoint X, we classify X with the class label k^* that maximizes the posterior probability of the class label given the data point:

$$k^* = \arg\max_k P(\text{class} = k|X)$$

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

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

1 Generative Models

There are two main types of models that can be used to train classifiers: **generative models** and **discriminative models**. Generative models involve explicitly forming:

1. A prior probability distribution over all classes $k \in \{1, 2, \dots, K\}$

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

2. A conditional probability distribution for each class k

$$f_k(X) = f(X|\text{class } k)$$

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 conclude (from Bayes' rule) that we are effectively solving

$$k^* = \arg\max_k P(\text{class} = k|X) = \arg\max_k \frac{P(k)\ f_k(X)}{f(X)} = \arg\max_k P(k)\ f_k(X)$$

In the case of the digit classification, we are solving for

$$k^* = \underset{k\in\{0,1,2...,9\}}{\arg\max}\ P(\text{digit} = k)\ f(\text{image}|\text{digit} = k)$$

2 QDA Classification

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

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

2.1 Estimating $f_k(.)$

For a particular class conditional probability distribution $f_k(.)$, if we do not have the true means and covariances μ_k, Σ_k, then our best bet is to estimate them empirically with the samples in our training data that are classified as class k:

$$\hat{\mu_k} = \frac{1}{n_k}\sum_{t_i=k} X_i$$
$$\hat{\Sigma_k} = \frac{1}{n_k}\sum_{t_i=k}(X_i - \hat{\mu_k})(X_i - \hat{\mu_k})^T$$

Note that the above formulas are not necessarily trivial and must be formally proven using MLE. Just to present a glimpse of the process, let's prove that these formulas hold for the case where we are dealing with 1-d data points. For notation purposes, assume that $\mathscr{D}_k = \{X_1, X_2, \ldots, X_{n_k}\}$ is the set of all training data points that belong to class k. Note that the data points are i.i.d. Our goal is to solve the following MLE problem:

$$\begin{aligned}
\hat{\mu_k}, \hat{\sigma_k}^2 &= \arg\max_{\mu_k, \sigma_k^2} P(X_1, X_2, ..., X_{n_k} | \mu_k, \sigma_k^2) \\
&= \arg\max_{\mu_k, \sigma_k^2} \ln(P(X_1, X_2, ..., X_{n_k} | \mu_k, \sigma_k^2)) \\
&= \arg\max_{\mu_k, \sigma_k^2} \sum_{i=1}^{n_k} \ln(P(X_i | \mu_k, \sigma_k^2)) \\
&= \arg\max_{\mu_k, \sigma_k^2} \sum_{i=1}^{n_k} -\frac{(X_i - \mu_k)^2}{2\sigma_k^2} - \ln(\sigma_k) - \frac{1}{2}\ln(2\pi) \\
&= \arg\min_{\mu_k, \sigma_k^2} \sum_{i=1}^{n_k} \frac{(X_i - \mu_k)^2}{2\sigma_k^2} + \ln(\sigma_k)
\end{aligned}$$

Note that the objective above is not jointly convex, so we cannot simply take derivatives and set them to 0! Instead, we decompose the minimization over σ_k^2 and μ_k into a nested optimization problem:

$$\min_{\mu_k, \sigma_k^2} \sum_{i=1}^{n_k} \frac{(X_i - \mu_k)^2}{2\sigma_k^2} + \ln(\sigma_k) = \min_{\sigma_k^2} \min_{\mu_k} \sum_{i=1}^{n_k} \frac{(X_i - \mu_k)^2}{2\sigma_k^2} + \ln(\sigma_k)$$

The optimization problem has been decomposed into an inner problem that optimizes for μ_k given a fixed σ_k^2, and an outer problem that optimizes for σ_k^2 given the optimal value $\hat{\mu_k}$. Let's first solve the inner optimization problem. Given a fixed σ_k^2, the objective is convex in μ_k, so we can simply take a partial derivative w.r.t μ_k and set it equal to 0:

$$\frac{\partial}{\partial \mu_k}\Big(\sum_{i=1}^{n_k} \frac{(X_i - \mu_k)^2}{2\sigma_k^2} + \ln(\sigma_k)\Big) = \sum_{i=1}^{n_k} \frac{-(X_i - \mu_k)}{\sigma_k^2} = 0 \implies \hat{\mu_k} = \frac{1}{n_k}\sum_{i=1}^{n_k} X_i$$

Having solved the inner optimization problem, we now have that

$$\min_{\sigma_k^2} \min_{\mu_k} \sum_{i=1}^{n_k} \frac{(X_i - \mu_k)^2}{2\sigma_k^2} + \ln(\sigma_k) = \min_{\sigma_k^2} \sum_{i=1}^{n_k} \frac{(X_i - \hat{\mu_k})^2}{2\sigma_k^2} + \ln(\sigma_k)$$

Note that this objective is not convex in σ_k, so we must instead find the critical point of the objective that minimizes the objective. Assuming that $\sigma_k \geq 0$, the critical points are:

- $\sigma_k = 0$: assuming that not all of the points X_i are equal to $\hat{\mu_k}$, there are two terms that are at odds with each other: a $1/\sigma_k^2$ term that blows off to ∞, and a $\ln(\sigma_k)$ term that blows off to $-\infty$ as $\sigma_k \to 0$. Note that the $1/\sigma_k^2$ term blows off at a faster rate, so we conclude that

  $$\lim_{\sigma_k \to 0} \sum_{i=1}^{n_k} \frac{(X_i - \hat{\mu_k})^2}{2\sigma_k^2} + \ln(\sigma_k) = \infty$$

- $\sigma_k = \infty$: this case does not lead to the solution, because it gives a maximum, not a minimum.

  $$\lim_{\sigma_k \to \infty} \sum_{i=1}^{n_k} \frac{(X_i - \hat{\mu_k})^2}{2\sigma_k^2} + \ln(\sigma_k) = \infty$$

- Points at which the derivative w.r.t σ is 0

$$\frac{\partial}{\partial\sigma}\Big(\sum_{i=1}^{n_k}\frac{(X_i-\hat{\mu}_k)^2}{2\sigma_k^2}+\ln(\sigma_k)\Big)=\sum_{i=1}^{n_k}-\frac{(X_i-\hat{\mu}_k)^2}{\sigma_k^3}+\frac{1}{\sigma_k}=0 \implies \hat{\sigma_k}^2=\frac{1}{n_k}\sum_{i=1}^{n_k}(X_i-\hat{\mu}_k)^2$$

We conclude that the optimal point is

$$\hat{\sigma_k}^2=\frac{1}{n_k}\sum_{i=1}^{n_k}(X_i-\hat{\mu_k})^2$$

2.2 QDA Optimization Formulation

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

$$\begin{aligned}
k^* &= \arg\max_k P(k)\ f_k(X) \\
&= \arg\max_k (\sqrt{2\pi})^d P(k)\ f_k(X) \\
&= \arg\max_k \ln(P(k))+\ln((\sqrt{2\pi})^d f_k(X)) \\
&= \arg\max_k \ln(P(k))-\frac{1}{2}(X-\hat{\mu_k})^T\hat{\Sigma_k}^{-1}(X-\hat{\mu_k})-\frac{1}{2}\ln(|\hat{\Sigma_k}|)=Q_k(X)
\end{aligned}$$

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

3 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 μ_k, but LDA is different from QDA in that it requires all of the distributions to share the same covariance matrix Σ. 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.

3.1 Estimating $f_k(.)$

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

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

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 X_i's by specifying a class k and the index within that class j:

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

4 LDA vs. QDA: Differences and Decision Boundaries

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 X-space. Given any two classes, the decision boundary represents the points in 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 X can be classified according to three cases:

$$k^* = \begin{cases} A & P(\text{class} = A|X) > P(\text{class} = B|X) \\ B & P(\text{class} = A|X) < P(\text{class} = B|X) \\ \text{Either } A \text{ or } B & P(\text{class} = A|X) = P(\text{class} = B|X) \end{cases}$$

The decision boundary is the set of all points in 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.

4.1 Identical Isotropic Gaussian Distributions

The simplest case is when the two classes are equally likely in prior, and their conditional probability distributions are isotropic with identical covariances. **Isotropic** Gaussian distributions have covariances of the form of $\Sigma = \sigma^2 I$, which means that their isocontours are circles. In this case, $f_A(.)$ and $f_B(.)$ have identical covariances of the form $\Sigma_A = \Sigma_B = \sigma^2 I$.

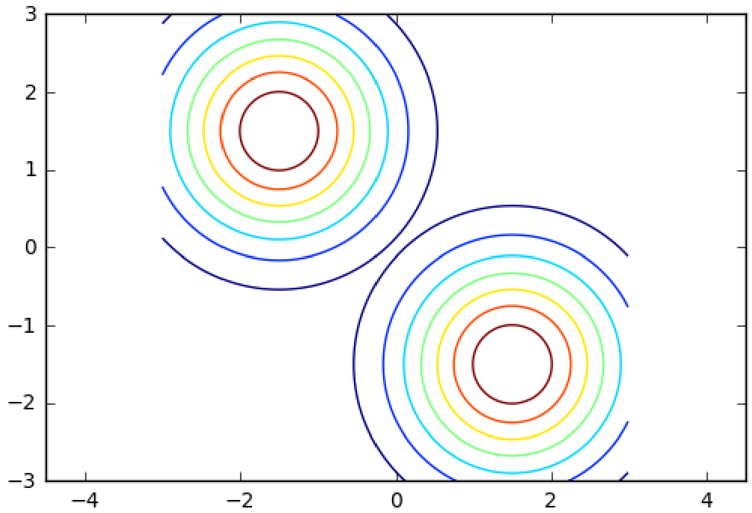

Figure 1: Contour plot of two isotropic, identically distributed Gaussians in $\mathbb{R}^2$. The circles are the level sets of the Gaussians.

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 $Q_k(X)$ from before, this can be expressed mathematically as:

$$\begin{aligned} Q_A(X) &= Q_B(X) \\ \ln\left(\frac{1}{2}\right) - \frac{1}{2}(X-\hat{\mu_A})^T\sigma^2 I(X-\hat{\mu_A}) - \frac{1}{2}\ln(|\sigma^2 I|) &= \ln\left(\frac{1}{2}\right) - \frac{1}{2}(X-\hat{\mu_B})^T\sigma^2 I(X-\hat{\mu_B}) - \frac{1}{2}\ln(|\sigma^2 I|) \\ (X-\hat{\mu_A})^T(X-\hat{\mu_A}) &= (X-\hat{\mu_B})^T(X-\hat{\mu_B}) \end{aligned}$$

The decision boundary is the set of points X for which $||X-\hat{\mu_A}||_2 = ||X-\hat{\mu_B}||_2$, which is simply the set of points that are equidistant from μ_A and μ_B. This decision boundary is linear because the set of points that are equidistant from μ_A and μ_B are simply the perpendicular bisector of the segment connecting μ_A and μ_B.

4.2 Identical Anisotropic Gaussian Distributions

The next case is when the two classes are equally likely in prior, and their conditional probability distributions are anisotropic with identical covariances. **Anisotropic** Gaussian distributions are simply all Gaussian distributions that are not isotopic.

In order to understand the difference, let's take a closer look at the covariance matrix Σ. Since Σ is a symmetric, positive semidefinite matrix, we can decompose it by the spectral theorem into $\Sigma = V\Lambda V^T$, where the columns of V form an orthonormal basis in $\mathbb{R}^d$, and Λ is a diagonal matrix with real, non-negative values. The entries of Λ dictate how elongated or shrunk the distribution is along each direction. To see why this is the case, let's consider a zero-mean Gaussian distribution $\mathcal{N}(0,\Sigma)$. We wish to find its **level set** $f(X) = k$, or simply the set of all points X such that the probability density $f(X)$ evaluates to a fixed constant k. This is equivalent to the level set $\ln(f(x)) = \ln(k)$, which further reduces to $x^T\Sigma^{-1}x = c$, for some constant c. Without loss of generality, assume that this constant is 1. The level set $x^T\Sigma^{-1}x = 1$ is an ellipsoid with axes $v_1, v_2, \dots, v_d$, with lengths $\sqrt{\lambda_1}, \sqrt{\lambda_2}, \dots, \sqrt{\lambda_d}$, respectively. Each axis of the ellipsoid is the vector $\sqrt{\lambda_i}v_i$, and we can verify that

$$(\sqrt{\lambda_i}v_i)^T\Sigma^{-1}(\sqrt{\lambda_i}v_i) = \lambda_i v_i^T\Sigma^{-1}v_i = \lambda_i v_i^T(\Sigma^{-1}v_i) = \lambda_i v_i^T(\lambda_i^{-1}v_i) = v_i^T v_i = 1$$

In the case of isotropic distributions, the entries of Λ are all identical, meaning the the axes of the ellipsoid form a circle. In the case of anisotropic distributions, the entries of Λ are not necessarily identical, meaning that the resulting ellipsoid may be elongated/shruken and also rotated.

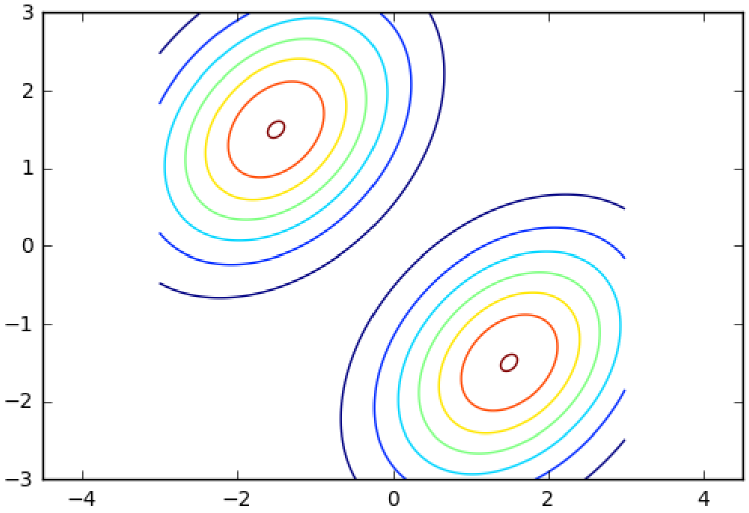

Figure 2: Two anisotropic, identically distributed Gaussians in $\mathbb{R}^2$. The ellipses are the level sets of the Gaussians.

The case when the two classes are identical anisotropic distributions can be reduced to the isotropic case simply by performing a change of coordinates that transforms the ellipses back into circles. Thus, the decision boundary is still linear.

4.3 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:

$$\begin{aligned}
Q_A(X) &= Q_B(X) \\
\ln(P(A)) - \frac{1}{2}(X-\hat{\mu_A})^T\hat{\Sigma}^{-1}(X-\hat{\mu_A}) - \frac{1}{2}\ln(|\hat{\Sigma}|) &= \ln(P(B)) - \frac{1}{2}(X-\hat{\mu_B})^T\hat{\Sigma}^{-1}(X-\hat{\mu_B}) - \frac{1}{2}\ln(|\hat{\Sigma}|) \\
\ln(P(A)) - \frac{1}{2}(X-\hat{\mu_A})^T\hat{\Sigma}^{-1}(X-\hat{\mu_A}) &= \ln(P(B)) - \frac{1}{2}(X-\hat{\mu_B})^T\hat{\Sigma}^{-1}(X-\hat{\mu_B}) \\
2\ln(P(A)) - X^T\hat{\Sigma}^{-1}X + 2X^T\hat{\Sigma}^{-1}\hat{\mu_A} - \hat{\mu_A}^T\hat{\mu_A} &= 2\ln(P(B)) - X^T\hat{\Sigma}^{-1}X + 2X^T\hat{\Sigma}^{-1}\hat{\mu_B} - \hat{\mu_B}^T\hat{\mu_B} \\
2\ln(P(A)) + 2X^T\hat{\Sigma}^{-1}\hat{\mu_A} - \hat{\mu_A}^T\hat{\mu_A} &= 2\ln(P(B)) + 2X^T\hat{\Sigma}^{-1}\hat{\mu_B} - \hat{\mu_B}^T\hat{\mu_B}
\end{aligned}$$

Simplifying, we have that

$$X^T(\hat{\Sigma}^{-1}(\hat{\mu_A}-\hat{\mu_B}))+\Big(\ln\big(\frac{P(A)}{P(B)}\big)-\frac{\hat{\mu_A}^T\hat{\mu_A}-\hat{\mu_B}^T\hat{\mu_B}}{2}\Big)=0$$

The decision boundary is the level set of a linear function $f(x)=w^Tx+k$. 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.

4.4 Nonidentical Distributions

This is certainly *not* the case in LDA. We have that:

$$\ln(P(A))-\frac{1}{2}(X-\hat{\mu_A})^T\hat{\Sigma_A}^{-1}(X-\hat{\mu_A})=\ln(P(B))-\frac{1}{2}(X-\hat{\mu_B})^T\hat{\Sigma_B}^{-1}(X-\hat{\mu_B})$$

Here, unlike the case when $\Sigma_A=\Sigma_B$, we *cannot* cancel out the quadratic terms in 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!

4.5 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:

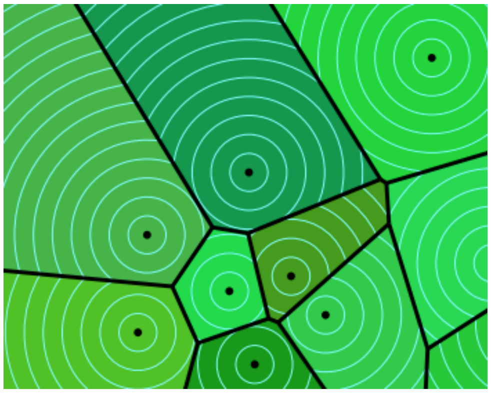

Figure 3: LDA: a collection of linear level set boundaries

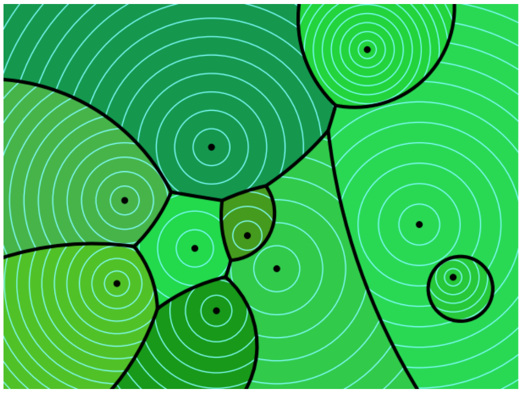

Figure 4: QDA: a collection of quadratic level set boundaries