CS 5785 – Applied Machine Learning – Lec. 6

Prof. Serge Belongie, Cornell Tech
Scribe: Luke Ahn, Ashley John.

Sept. 17, 2018

1 Multivariate Gaussian/Normal Density

One of the most commonly used pdfs in engineering is the multivariate Gaussian or normal density:

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^\top \Sigma_k^{-1}(x-\mu_k)} \tag{1}$$

The constant term $\frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}}$ ensures that $f_k(x)$ integrates to 1. For additional intuition, plug in $p = 1$ and see that it reduces to the familiar *univariate* (or 1D) Gaussian density, parameterized by a mean μ and variance σ^2. The equation (1) will come down to:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2}$$

In particular, we often use multivariate normals as *class conditional densities*, e.g., to describe the height and weight of a pet given that it is a dog. We say "X conditioned on class k is normally distributed" and write:

$$X|G = k \sim \mathcal{N}(\mu_k, \Sigma_k) \tag{3}$$

In this expression, μ_k and Σ_k represent the *mean vector* and *covariance matrix* of the density for the kth class, respectively, and p is the number of dimensions of the feature vector x, as usual. The covariance matrix is sometimes called the variance-covariance matrix since the variances of the individual variables are on the diagonal and covariances are off-diagonal.

Figures 1 and 2 show two ways of visualizing this distribution for $p = 2$, a special case known as a *bivariate normal.* Figure 1 highlights that the fact that the level sets of a bivariate Gaussian are ellipses. Figure 2 shows three examples of bivariate Gaussians as surface plots and *heatmaps.* If the covariance matrix is represented by an identity matrix, the contour plots of probability density will be concentric circles. In any other case, looking at covariance matrices doesn't make any sense, you can't predict it based on the values. Eigenvalues and eigenvectors should be used instead.

Figure 3 shows an example of bivariate Gaussians used to model features of two birds (sandpiper vs. hawk). In this example, we have two class conditional densities with different means but equal covariance matrices, and the features

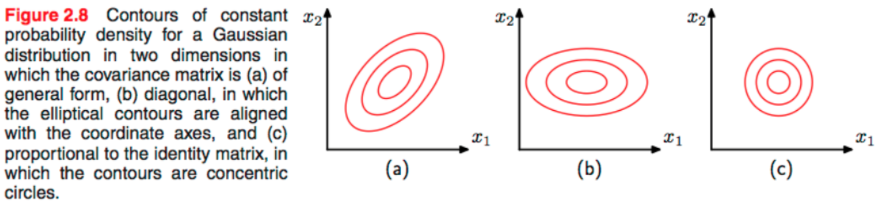

Figure 1: [Bishop]

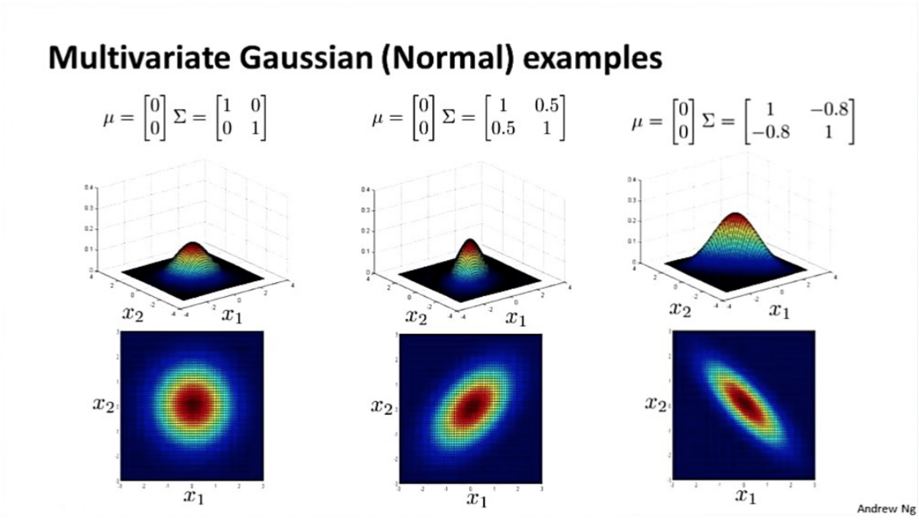

Figure 2: [A. Ng]

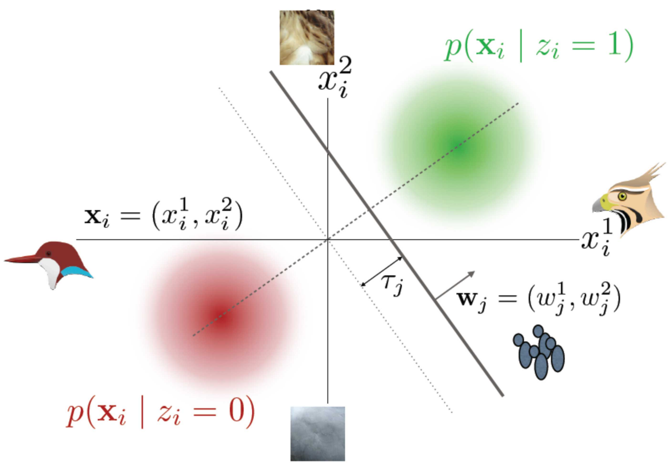

Figure 3: [Welinder]

capture characteristics of the beak and color of the plumage. The figure also foreshadows how we can use these densities to define a decision boundary.

One practical limitation of the Gaussian is that it is *unimodal*, which is another way of saying that the density just has one bump. Such a density isn't a good fit for classes comprising distinct subclasses, e.g., the color of bell peppers. Multimodal densities such as a *Gaussian Mixture Models*, which we'll discuss later in the class, can address this problem.

1.1 Quadratic Forms

Look at the exponent in Equation 1. Assuming $\mu_k = 0$, which we refer to as the *mean zero* or *centered* case, $(x - \mu_k)^\top \Sigma_k^{-1}(x - \mu_k)$ has the form $f(x) = x^\top Qx$, which is known in linear algebra as a *quadratic form*. Looking at the dimensions of these terms, the exponent resolves to a scalar. The entries in the matrix Q control the shape of the function $f(x)$, as illustrated in Figure 1. An ellipse in that figure represents the locus of values x such that $f(x) = x^\top Qx =$ some constant. In the case of Equation 1, the covariance matrix Σ_k^{-1} is *positive definite*, a property that we will discuss later. People often think of covariance matrices as ellipses: perfect circles mean no correlation.

This quadratic form is of interest to us since we often work with log likelihoods instead of raw likelihoods. Returning to Equation 1, when you take its log, you get the quadratic form discussed above.

2 Linear Discriminant Analysis

Linear Discriminant Analysis[1] is a supervised classification method that assumes the class conditional densities are Gaussian and that all of the covariance matrices are equal, i.e., $\Sigma_k = \Sigma$. It can also be used for dimensionality reduction, which takes in very large feature vectors and collapses them into a small number of meaningful dimensions. To see how it works, let π_k denote the prior probability of class k (which means $\sum_{k=1}^K \pi_k = 1$) and apply Bayes' theorem to express the posterior probability as follows:

$$Pr(G = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^K f_l(x)\pi_l}$$

Recall that $f_k(x)$ is the class-conditional density and the denominator is the normalization factor or marginal likelihood. Bayes' rule allows us to take the class conditional density and turn it around so we can predict the class given our data. Assume for now that π_k, μ_k and Σ_k are all known. Now look at the log-odds we get by comparing two classes k and l:

$$\log \frac{Pr(G = k|X = x)}{Pr(G = l|X = x)} = \log \frac{f_k(x)}{f_l(x)} + \log \frac{\pi_k}{\pi_l}$$

Notice how the log-odds on the priors appears as a constant offset or shift. Assuming Gaussian class conditional densities, we get

$$\log \frac{Pr(G = k|X = x)}{Pr(G = l|X = x)} = \log \frac{\pi_k}{\pi_l} - \frac{1}{2}(\mu_k + \mu_l)^\top \Sigma^{-1}(\mu_k - \mu_l) + x^\top \Sigma^{-1}(\mu_k - \mu_l)$$

[1]Unfortunately the initialism LDA is now taken by Latent Dirichlet Allocation, so we can't use it here.

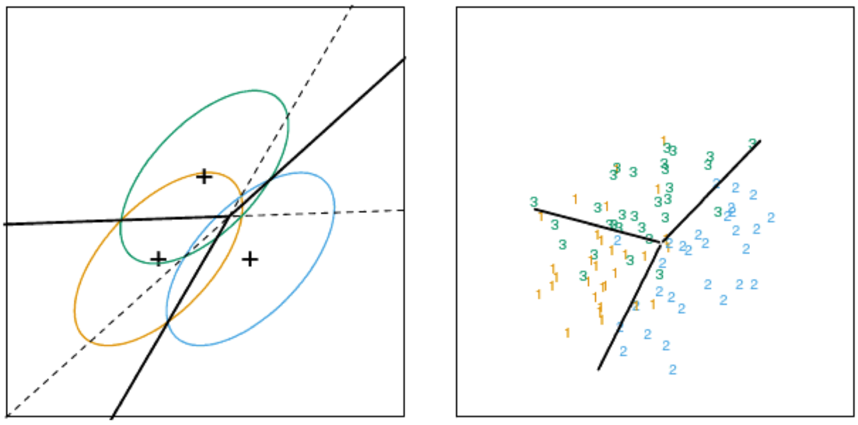

Figure 4: Decision boundaries between 3 classes modeled using bivariate normals with equal covariance.[HTF Fig. 4.5]

Note that the normalizations constants canceled out and, since all the Σ_ks are equal, we don't have any terms quadratic in x. In other words, x only appears linearly in this expression, which has the form $\alpha_{k0} + \alpha_k^\top x$.

The locus where $Pr(G = k|X = x) = Pr(G = l|X = x)$ is a line (plane, hyperplane, etc.) and this is the *decision boundary*: on one side, class k is more likely than class l, vice versa on the other side.

Figure 4 shows the decision boundaries between three classes in 2 dimensions. If the covariances weren't equal, the decision boundaries wouldn't be straight. In such cases one can use Quadratic Discriminant Analysis, which we won't be covering.

2.1 Linear Discriminant Function

Formally speaking, the *linear discriminant functions* have the form

$$\delta_k(x) = x^\top \Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^\top \Sigma^{-1}\mu_k + \log \pi_k$$

with decision rule:

$$G(x) = \arg\max_k \delta_k(x)$$

Assuming we've already calculated the mean and covariance and we've estimated the prior, to use this as a classifier we take an observation x, plug this into discriminant function for each class and see which one yields the maximum discriminant function value.

If the covariance matrices are spherical (i.e., $\Sigma \propto I$) and the priors are equal (e.g., sandpipers just as likely as hawks at a campsite), the decision boundary is simply the perpendicular bisector of the line joining the two means. Changing the prior (or class balance) has the effect of shifting the decision boundary along that line. In the case of non-spherical covariance matrices, as in Fig. 4, the decision boundary still bisects the line joining the means but is no longer perpendicular.

2.2 Parameter Estimation

We've been assuming that the priors, mean and covariance were estimated for us thus far. Here's how you estimate them.

Priors:

$$\hat{\pi}_k = \frac{N_k}{N}$$

where N_k is the number of class k observations and N is the total number of observations.

Sample mean:

$$\hat{\mu}_k = \frac{1}{N_k} \sum_{g_i = k} x_i$$

Here the sum is over the feature vectors belonging to class k.

Sample covariance:

$$\hat{\Sigma} = \frac{1}{N-K} \sum_{k=1}^{K} \sum_{g_i = k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^\top$$

Covariance matrices are also known as *centered second moment* matrices. The centering operation refers to subtracting the class mean. The second moments refer to the sum of the outer products of the feature vectors with themselves. You can think of this as the vector counterpart to squaring for scalars. Each term in that sum is a $p \times p$ matrix containing all the possible products between pairs of entries in a feature vector. The outer summation is over all K classes; we can do this *pooling* operation because we assumed that all classes have same covariance matrix.

2.3 Comparison with Logistic Regression

The log posterior odds in Linear Discriminant Analysis has the form

$$\alpha_{k0} + \alpha_k^\top x$$

which looks an awful lot like

$$\beta_{k0} + \beta_k^\top x$$

from logistic regression. The difference lies in how the coefficients are estimated. The bottom line is that the results of using the two approaches are often very similar. LR is a discriminative classifier. LR has the advantage of making fewer assumptions about how the data are distributed. Linear Discriminant Analysis is a generative classifier, and it has the advantage of offering low dimensional projections of the data, useful for visualization. Linear Discriminant Analysis assumes a multivariate Gaussian distribution, so it can generate additional data as needed.

We'll discuss this in the next lecture.