Statistical Machine Learning
Part 5

Horacio Gómez-Acevedo
Department of Biomedical Informatics

March 9, 2021

Probability refresher

Formal definition. A probability space is a triplet $(A, \mathfrak{A}, P)$, where A is a (non-empty) set, $\mathfrak{A}$ contains subsets of A called *events* (those events form a σ-algebra of A) and a function $P\colon \mathfrak{A} \to [0,1]$ (the so-called *probability function*).
Example. Let's consider the experiment of rolling a die. A will be the outcomes $A = \{1, 2, \ldots, 6\}$ and $\mathfrak{A}$ will be the subsets of A. For instance, the event E consisting of the even number output is $\{2, 4, 6\}$. The probability function will be defined as

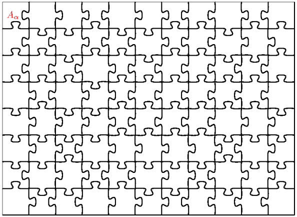
$$P(E) = \frac{\text{number of elements of } E}{\text{number of elements in } A} = \frac{3}{6}$$

Complete system of events

We called a *complete system of events* a collection $\{A_\alpha\}_{\alpha \in I}$ that satisfy:

- $A_\alpha \cap A_\beta = \emptyset$
- $\cup_{\alpha \in I} A_\alpha = A.$

A jigsaw puzzle is a simple visualization of this. A is the whole rectangle and each piece is one of the A_α.



Conditional Probabilities

If B and C are events and $P(C) > 0$, we say that the probability that the event B has occurred given that C has occurred is given by

$$P(B|C) = \frac{P(B \cap C)}{P(C)}$$

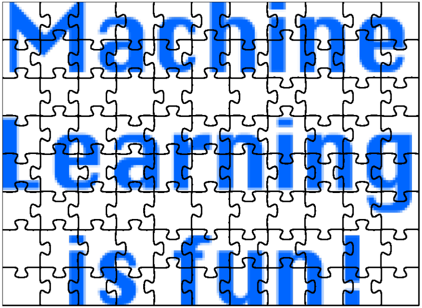
Total Probability Theorem. If we have a complete system of events $\{A_\alpha\}_\alpha \in I$ with $P(A_\alpha) > 0$ for each $\alpha \in I$, then for an arbitrary event E it holds

$$P(E) = \sum_\alpha P(E|A_\alpha)P(A_\alpha)$$

It is clear that every term $P(E|A_\alpha)P(A_\alpha) = P(A_\alpha \cap E)$. Thus, the probability of E is the sum of every part of E in each A_α.

Total Probability Theorem

Suppose our space is again the jigsaw puzzle rectangle and each piece is A_α as before. The probability measure will be the area cover by an event. In this case our "event" represents the blue letters. Visually the probability of the whole event will be the sum of the blue area in each of the pieces, thus the name "total probability"



Bayes Theorem

Following the same framework as before, we have one of the most useful results in probability
Bayes theorem. If E is an event with $P(E) > 0$, then

$$P(A_\beta|E) = \frac{P(E|A_\beta)P(A_\beta)}{\sum_{\alpha \in I} P(E|A_\alpha)P(A_\alpha)} \tag{1}$$

Note that the numerator of (1) is $P(E \cap A_\beta)$ whereas the denominator is $P(E)$ according to the total probability theorem!

Bayesian Perspective

Unlike the traditional (frequentist) inference, the Bayesian approach tries to determine conclusions about a parameter θ or unobserved data $\overline{y}$ as probability statements. For example, given the observed data y, Bayesian statements are based on the conditionals such as $P(\overline{y}|y)$ for *new data* $\overline{y}$, or $P(\theta|y)$ for a parameter θ. So following our Bayes theorem,

$$P(\theta|y) = \frac{P(\theta)P(y|\theta)}{P(y)} \tag{2}$$

where $P(y) = \sum_\theta P(\theta)P(y|\theta)$ (or in the continuous case $P(y) = \int P(\theta)P(y|\theta)d\theta$).

The expression $P(\theta)$ is called the **prior**, whereas $P(\theta|y)$ is called **posterior**, and $P(y|\theta)$ is called **likelihood**, and $P(y)$ is called **evidence**.

Note that in (2) the denominator does not depend on θ. So, the unnormalized posterior

$$P(\theta|y) \propto P(\theta) \cdot P(y|\theta)$$

Density functions

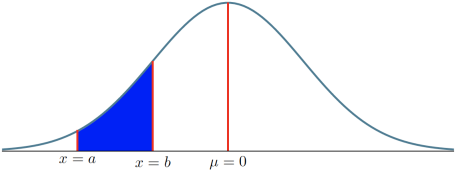
When we have a continuous random variable X, the probability is calculated by integrals instead of sums. The **probability density function** $f(x)$ describes the probability of an event.

For instance, let's suppose we have a random variable X distributed as a $N(0,1)$. The density function in this case is defined as

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$$

Then, the probability that our random variable lies in the interval $[a,b]$ is given by

$$P(a \leq X \leq b) = \int_a^b f(x)dx = \int_a^b \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)dx$$



Linear Discriminant Analysis

We need to classify an observation into one of K distinct classes ($K \geq 2$). The response variable Y can take on K possibilities. Let π_k denote the *prior* probability that a randomly chose observation comes from the kth class. Let $f_k(X) = P(X = x|Y = k)$ denote the *density function* of X for an observation that comes from the kth class. We know by the Bayes' theorem

$$P(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{i=1}^{K} \pi_i f_i(x)} \tag{3}$$

Let's simplify *posterior probability* , so $p_k(X) = P(Y = k|X)$
What have we achieved thus far?

- We have a classifier based on posterior probabilities (3) by selecting the class with the highest probability.
- The prior information can easily be achieved
- The terms f_k are hard to calculate.

Linear Discriminant Analysis $p = 1$

Let's consider the case that we have K classes and all our observations are in the real line $\mathbb{R}$.

Case 1. $f_k(x)$ **are normal densities with common variance** σ^2**.** In this case

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2,\right)$$

where μ_k are the (estimates) of the mean for each class. Note that the priors π_k are estimated based on the proportion of samples in class k from the training set. Thus,

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{i=1}^{K} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_i)^2\right)}$$

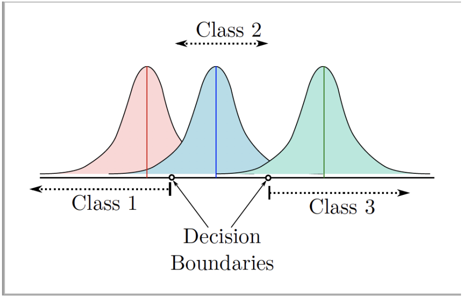
The classification rule for $X = x$ states that we will select the class with the largest posterior probability.

Discriminant Function for Case 1.

It can be shown that

$$\begin{aligned}\text{Class of } x = &\operatorname{argmax}_k\{p_k(x)\} = \operatorname{argmax}_k \delta_k(x) \\ &= \operatorname{argmax}_k x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)\end{aligned} \tag{4}$$

where argmax returns the index of the largest value of $p_k(x)$. Thus, the discriminant (4) is linear ($ax + b$) and there are $K - 1$ points that split the line into K classes.



Linear Discriminant Analysis (Case 1)

Even if we assume that each class has a normal distribution, we need to estimate $\mu_1, \ldots, \mu_K$, $\pi_1, \ldots, \pi_K$ and σ.
Formula (4) will still be valid if we plug in the following estimates

$$\hat{\mu}_k = \frac{1}{n_K} \sum_{i: y_i \text{ in class } k} x_i$$

$$\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i: y_i \text{ in class } k} (x_i - \hat{\mu}_k)^2$$

$$\hat{\pi}_k = \frac{n_k}{n}$$

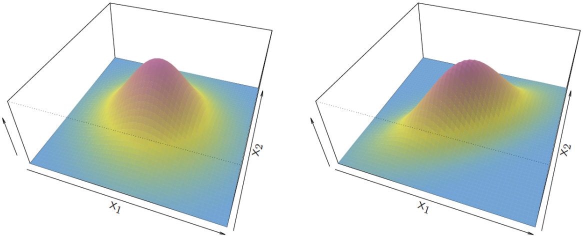
Linear Discriminant Analysis (Case 2)

Once we have found our discriminant function for the case $p = 1$, now we will move to higher dimensions ($p \geq 2$). Let's assume that $\mathbf{X} = (X_1, \ldots, X_p)$ are drawn from a multivariate Gaussian with a class-specific mean vector $E(\boldsymbol{X}) = \boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)$,and a common variance-covariance matrix $\boldsymbol{\Sigma}$. The general multivariate normal density is written as

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right) \quad (5)$$

where $|\boldsymbol{\Sigma}|$ is represents the determinant of the matrix $\boldsymbol{\Sigma}$. So, the LDA for higher dimensions assumes that observations in the kth class are drawn from a multivariate distributions $N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$

Linear Discriminant Analysis (Case 2)



As for the discriminant function for higher dimensions

$$\delta_k(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k + \log(\pi_k)$$

Linear Discriminant Analysis (Case 2)

The Bayes decision boundaries will be calculated by pairs of indexes. Namely, if $\boldsymbol{x}$ is in the decision boundary of k and l then

$$\boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \boldsymbol{\mu}_k \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k = \boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_l - \frac{1}{2} \boldsymbol{\mu}_l \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_l$$

Solving this equation reduces to something like this

$$\boldsymbol{x}^T \cdot \boldsymbol{v} = r$$

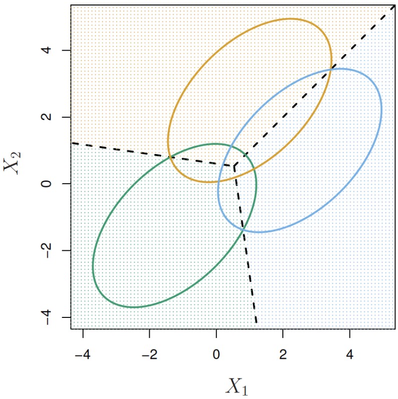
Thus, we need to calculate the vector $\boldsymbol{v} = (v_1, \ldots, v_p)$ such that

$$v_1 x_1 + \cdots + v_p x_p = r$$

In $\mathbb{R}^2$ this equation represents a line.

Linear Discriminant Analysis (Case 2)

Thus, when we have multiple classes there will be lines dividing the plane as shown below



Linear Discriminant Analysis

If we use LDA for binary classification we observe

- LDA has very low classification errors, which is due to the fact that approximates the Bayes classification.
- LDA will be off when the confusion matrix is "unbalanced". One remedy for this situation is to lower the threshold of 50% in the classifier to say 25%.

Quadratic Discriminant Analysis

We extend one of our main assumptions from the LDA as follows

- The k-th class has its own a variance-covariance matrix $\boldsymbol{\Sigma}_k$ (not the common one $\boldsymbol{\Sigma}$ as in LDA)

This apparently minimal changes has an impact on the discriminant equation

$$\delta_k(\boldsymbol{x}) = -\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\log|\boldsymbol{\Sigma}_k| + \log(\pi_k) \quad (6)$$

Note that the part on red makes the discriminant function quadratic! So a new observation

$$\text{Class of } \boldsymbol{x} = \text{argmax}_k \delta_k(\boldsymbol{x})$$

QDA or LDA

The estimation of the variance-covariance matrix in $\mathbb{R}^p$ requires $p(p+1)/2$ (essentially all the off diagonal terms), the calculation of the K variance-covariance matrices requires $Kp(p+1)/2$.

- LDA is much less flexible classifier than QDA. Thus it has lower variance.
- In LDA the assumption of a shared variance-covariance matrix can be terribly wrong. In this case, QDA is in order.
- For "small" sample sizes LDA is preferred as it reduces the variance. But as the sample size increases QDA should be used.

References

Materials and some of the pictures are from (1),(2), and (3).

1. Gareth James et al. *An Introduction to Statistical Learning with applications in R*. Springer (2015)
2. Richard O. Duda et al. *Pattern Classification* John Wiley (2001).
3. Aurélien Géron. *Hands-on Machine Learning with Scikit-Learn & TensorFlow* O'Relly (2017)
4. Wiebe R. Pestman *Mathematical Statistics* de Gruyter (1998)

I have used some of the graphs by hacking TiKz code from StakExchange, Inkscape for more aesthetic plots and other old tricks of TEX