**Optimal Bayes Classifier**

* **Supervised Learning**

This notebook summarises the theory and the derivation of the optimal bayes classifier. It then provides a comparison of the boundaries of the Optimal and Naive Bayes classifiers.

**Background and theory**

Consider now that instead of having a continuous output variable Y, we have instead a categorical output variable G. This model is summarized as

* **Input**: X∈Rp comes from a pp dimensional space
* **Output classification** G∈G where G is a random variable corresponding to the discrete output value, and GG is the discrete output space.
* **Joint distribution** on the input and output Pr(X,G)=[(x1,g1),(x1,g2)...(xm,gm)]
* **Goal** is to learn a function f(x):Rp→Gf which takes inputs from the pp dimensional input space and maps them to the discrete output space

A first step is to decide on an appropriate loss function, as the usual \qq{squared loss} is not appropriate for discrete outputs. Instead we will use the simple \qq{0-1 loss} function which is defined as follows.

Define the loss as a K×K matrix, where K=card(G) where the matrix will have 0 on the diagonal and non-negative values otherwise. So the loss L(k,l) is the k,l entry of the matrix, and is the cost of classifying kk as ll. For example, in the case of 3 classes we could get

⎣011

101

110⎦

Which means we can write the 0 - 1 loss function as:

L(k,l) = 0 if k=l

1 if k≠l

L(k,l) = I(k≠l)=1−I(k=l)

The Expected Predicted Error (EPE) is therefore:

EPE(f^(x))=E[L(G,f^(X))]

Where the expectation is taken with respect to the joint distribution Pr(X,G). Again we can condition on XX to obtain

EPE(f^(x))=EXEG|X[L(G,f^(X))|X]

EPE(f^(x))=EX∑k=1KL[k,f^(X)]Pr(k|X)

Where k=1,...,K are all the possible values that the random variable GG can take, i.e. the set GG. Note that this is the discrete version which is analogous to the derivations discussed in the previous section.

As we want to minimize the expected loss we can do the following:

f^(x) = argming∑k=1KL[k,g]Pr(k|X)

= argming∑k=1K(1−I(k=g))Pr(k|X)

=argmaxg∑k=1KI(k=g)Pr(k|X)

Since the indicator function is 1 when k=g we get

f^(x) = argmaxgPr(g|X=x)=MAP

In other words, the optimal Bayes decision rule is to choose the class presenting the maximum posterior probability, given the particular observation at hand. Classifiers such as these are called **Bayes Optimal Classifier** or **Maximum a Posteriori** classifiers.

Since, for a given observation xx, the marginal distribution of p(x)p(x) is constant in the denominator of Bayes theorem, we can simplify this decision rule further as:

f^(x) = argmaxgPr(g|X=x)

= argmaxgPr(x|g)p(g)p(x)

= argmaxgPr(x|g)p(g)

= argmaxglogPr(x|g)+logp(g)

This form makes clear that the MAP decision rule tries to reach a compromise between the a priori expectations p(g) and the evidence provided by the data via the likelihood function p(x|g).

**Optimal Bayes Classifier**

The Optimal Bayes classifier chooses the class that has greatest a posteriori probability of occurrence (so called **maximum a posteriori estimation, or MAP**). It can be shown that of all classifiers, the Optimal Bayes classifier is the one that will have the lowest probability of miss classifying an observation, i.e. the lowest probability of error. So if we know the posterior distribution, then using the Bayes classifier is as good as it gets.

In real-life we usually do not know the posterior distribution, but rather we estimate it. The **Naive Bayes classifier** approximates the optimal Bayes classifier by looking at the empirical distribution and by assuming independence of predictors. So the Naive Bayes classifier is not itself optimal, but it approximates the optimal solution.