Additional Information of chapter2

Ken Wang

As we know, for regression problem, the best decision is E{y|x}, so let us rethink of generalization error and this decision to infer a more intuitive understanding to reveal the element of regression.

Let us recall the generalization error.

$$E(L) = \iint L(f(x), y)\, p(x, y)\, dxdy$$

And, For square loss, we have

$\{f(x) - y\}^2 = \{f(x) - E(y|x) + E(y|x) - y\}^2$

$= \{f(x) - E(y|x)\}^2 + 2\{f(x) - E(y|x)\}\{E(y|x) - y\} + \{E(y|x) - y\}^2$

Let us re-calculate the generalization error,

$$E(L) = \int \{f(x) - E(y|x)\}^2 p(x)\, dx + \int var[y|x] p(x)\, dx$$

When the first term, f(x)=E(y|x), it will take minimize solution, meanwhile, the second term represent the conditional variance of response variable give by x. It represent the inner variant of the data(noise), because it has no relationship with f(x), it represent the part which can't be minimized in loss function.

Inference and decision

We can separate the problems of classification into two stage: '**inference**' and '**decision**'. The inference stage in which we use training data to learn a model for $p(C_k|\mathbf{x})$. and the subsequent decision stage in which we use these posterior probabilities to make optimal class assignments. An alternative possibility would be to solve both problems together and simply learn a function that maps inputs x directly into decisions. Such a function is called a discriminant function.

In fact, we can identify three distinct approaches to solving decision problems, all of which have been used in practical applications. These are given, in decreasing order of complexity, by

(a) First solve the inference problem of determining the class-conditional densities $p(\mathbf{x}|C_k)$ for each class C_k individually. Also separately infer the prior class probabilities $p(C_k)$. Then use Bayes' theorem in the form.

$$p(C_k|x) = \frac{p(x|C_k)\, p(C_k)}{p(x)}$$

Equivalently, we can model the joint distribution $p(\mathrm{x}, C_k)$ directly and then normalize to obtain the posterior probabilities. Having found the posterior probabilities, we use decision theory to determine class membership for each new input **x**. Approaches that explicitly or implicitly model the distribution of inputs as well as outputs are known as **generative models**, because by sampling from them it is possible to generate synthetic data points in the input space. **Example**. Gaussian mixed model

(b) First solve the inference problem of determining the posterior class probabilities $p(C_k|\mathbf{x})$, and then subsequently use decision theory to assign each new **x** to one of

the classes. Approaches that model the posterior probabilities directly are called **discriminative models. Example.** Logistic regression model

(c) Find a function $f(\mathbf{x})$, called a discriminant function, which maps each input $\mathbf{x}$ directly onto a class label. For instance, in the case of two-class problems, $f(\cdot)$ might be binary valued and such that $f = 0$ represents class C_1 and $f = 1$ represents class C_2. In this case, probabilities play no role. **Example.** Support Vector Machine.

Advantage and disadvantage

Approach (a) is the most demanding because it involves finding the joint distribution over both $\mathbf{x}$ and C_k. For many applications, $\mathbf{x}$ will have high dimensionality, and consequently we may need a large training set in order to be able to determine the class-conditional densities to reasonable accuracy. However, we can infer the marginal distribution of x from Bayes' theorem. This can be useful for detecting new data points that have low probability under the model and for which the predictions may be of low accuracy, which is known as **outlier detection** or **novelty detection.**

However, if we only wish to make classification decisions, then it can be wasteful of computational resources, and excessively demanding of data, to find the joint distribution $p(\mathbf{x}, C_k)$ when in fact we only really need the posterior probabilities $p(C_k|\mathbf{x})$, which can be obtained directly through approach (b).

Combining Problem

For complex applications, we may wish to break the problem into a number of smaller subproblems each of which can be tackled by a separate module. For example, in our hypothetical medical diagnosis problem, we may have information available from, say, blood tests as well as X-ray images. Rather than combine all of this heterogeneous information into one huge input space, it may be more effective to build one system to interpret the Xray images and a different one to interpret the blood data. As long as each of the two models gives posterior probabilities for the classes, we can combine the outputs systematically using the rules of probability. One simple way to do this is to assume that, for each class separately, the distributions of inputs for the X-ray images, denoted by $\mathbf{x}_I$, and the blood data, denoted by $\mathbf{x}_B$, are independent, so that

$$p(x_I, x_B|C_k) = p(x_I|C_k)\, p(x_B|C_k)$$

This is an example of **conditional independence** property, because the independence holds when the distribution is conditioned on the class C_k. The posterior probability, given both the X-ray and blood data, is then given by

$$p(C_k|x_I, x_B) \propto p(x_I, x_B|C_k)\, p(C_k) \propto \frac{p(C_k|x_I)\, p(C_k|x_B)}{p(C_k)}$$

Thus we need the class prior probabilities $p(C_k)$, which we can easily estimate from the fractions of data points in each class, and then we need to normalize the resulting posterior probabilities so they sum to one. This typical conditional independent assumption is an example of Naïve Bayesian Classification. And we will see another factorization in VBI. Also, we will have better understanding of conditional independent assumption after the introduction of bayesian network in chapter.Graph theory.