Decision theory and probability model

Ken Wang

1. Probability theory

Consider two random variables X and Y, X has M different valus,$\{x_{1,\dots,}x_M\}$, Y has N different values, $\{y_1 \dots y_L\}$. And I use x_i ,x_j to represent the specific values of two variables. Now suppose I have N times sampling experiments, and use n_{ij} to represent in a experiments, we have $x=x_i$ and $y=y_j$.

1.1 Distribution

Assume we annot the probability of $x=x_i$, $y=y_j$ is $p(x=x_i, y=y_j)$, so it equal to:

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N} \tag{1}$$

Here we are implicitly considering the limit $N\rightarrow\infty$. Also, if we want to calculate the probability of $x=x_i$, we can summarize the times of x_i presents. Assume it equals to c_i

$$p(X = x_i) = \frac{c_i}{N} \tag{2}$$

Because c_i is the sum of ith columne of "sampling matrix", so we have,

$$p(X = x_i) = \frac{c_i}{N} = \frac{\sum_{j=1}^{L} n_{ij}}{N} = \sum_{j=1}^{L} p(X = x_i, Y = y_j) \tag{3}$$

Here, We get the **sum rule** of probability theory, and we call $p(X=x_i)$ is the **marginal distribution.**

If we consider only those instances for which X = xi, then the fraction of such instances for which Y = yj is written p(Y = yj|X = xi) and is called the conditional probability of Y = yj given X = xi. It is obtained by finding the fraction of those points in column i that fall in cell i,j and hence is given by.

$$p(Y = y_j|X = x_i) = \frac{n_{ij}}{c_i} \tag{4}$$

Combine the equations (1), (2), (4), we have

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{c_i}\frac{c_i}{N} = p(Y = y_j|X = x_i)\, p(X = x_i) \tag{5}$$

Here, We get the product rule of probability theory.

Sum rule $$p(X) = \sum_{Y} p(X, Y) = \int_{Y} p(X, Y)$$

Product rule $$p(X, Y) = p(Y|X)\, p(X)$$

From the product rule, together with the symmetry property p(X, Y) = p(Y, X), we immediately obtain the following relationship between conditional probabilities.

$$p(Y|X) = \frac{p(X|Y)\, p(Y)}{p(X)}$$

Which is called **Bayes'theorem**

From the definition of probability, these fractions would equal the corresponding probabilities p(Y) in the limit $N \to \infty$. We can view the histogram as a simple way to model a probability distribution given only a finite number of points drawn from that distribution. **Modelling distributions from data lies at the heart of statistical pattern recognition.**

Example1. **The fruit selection.**

Prior: the probability available before we observe.

Posterior: it is the probability obtained after we have observed.

Consider another situation, if X, Y is independent, we have:

$$p(X, Y) = p(X)\, p(Y)$$

So, the condition probability of Y given X is equal to its marginal distribution

2. Density function

As well as considering probabilities defined over discrete sets of events, we also wish to consider probabilities with respect to continuous variables. We shall limit ourselves to a relatively informal discussion. If the probability of a real-valued variable x falling in the interval (x, x + δ x) is given by p(x) δ x for δ x $\to$ 0, then p(x) is called the probability density over x. The probability that x will lie in an interval (a, b) is then given by

$$p(x \in (a, b)) = \int_a^b p(x)\, dx$$

It need to satisfy two conditions

$$\int_{-\infty}^{+\infty} p(x) = 1$$

$$p(x) \geqslant 0$$

And, the sum rule and product rule also can be applied on density function

3. Expectation and covariance

One of the most important operations involving probabilities is that of finding weighted averages of functions. The average value of some function f(x) under a probability distribution p(x) is called the expectation of f(x) and will be denoted by E[f]. For a discrete distribution, it is given by:

$$E[f] = \sum_x p(x)\, f(x)$$

so that the average is weighted by the relative probabilities of the different values of x. In the case of continuous variables, expectations are expressed in terms of an integration with respect to the corresponding probability density.

$$E[f] = \int_x p(x)\, f(x)$$

We can also consider a conditional expectation with respect to a conditional distribution, so that

$$E[f|y] = \int_x p(x|y) f(x)$$

The variance of f(x) can be defined as

$$var[f] = E[(f - E[f])^2]$$

We also can define the covariance for two random variable

$$cov(x, y) = E_{x,y}[\{x - E[x]\}\{y - E[y]\}] = E_{x,y}[xy] - E[x]E[y]$$

In the condition of two random variable vector, it is the co-variance matrix

4. Bayesian and Frequency

4.1 Bayesian

The points of views of Bayesian is start from two works, the Bayesian'therom and Bayesian hypothesis, In such circumstances, we would like to be able to quantify our expression of uncertainty and make precise revisions of uncertainty in the light of new evidence, as well as subsequently to be able to take optimal actions or decisions as a consequence.

$$p(\Theta|D) = \frac{p(D|\Theta)\, p(\Theta)}{p(D)}$$

The point is that we consider the parameters as the random variale. So actually we are modeling the parameters distribution. So we have

$$posterior \propto likelihood \times prior$$

The above distributions are actually the function of parameters. So in the perspective of bayesian, the error of estimate is from the distribution of parameters, because we only a single data set D.

Overview of Bayesian inference:

1). Considering the unknown parameters as random variables(or random vector), annot it is theta. So the joint distribution of samples will be the condition distribution of parameters.

2). We need to dertermine the distribution of prior, it is based on the knowledge of parameters.**This is the most controversial step in the Bayesian approach.**

3). Using the condition density and prior, we can get the joint distribution between data and parameters, and the distribution of data(normalized factor). So that we can get the posterior distribution.

4). Using the posterior density distribution we can make inference of parameters.

In the step2, if we don't have any prior information to help us to determine the prior, Bayesian can propose use uniform distribution as prior, the philosophy to determine the prior is called Bayesian hypothesis. So the criticisms of Bayesian is mainly focus on two points:

1. Is it reasonable to consider parameters as random variables?

2. Dose prior distribution exist? How to choose?
(Emperical Bayesian, the most attractive method in Bayesian, base on the data to determine the prior distribution)

4.2 frequentist

Frequency, also we call **classical, sampling.** It is start from Pearson K. developed by R.A. Fisher(my idol). and finishsed the theoritical proof by Neyman. J.
It is including the point estimate, hypothesis test, maximum likelihood and OLS, confidence interval estimate.....The frequncy family's idea is that parameter is not random variable, it's fixed, **so the error of estimate is from the data!** This is why it is called frequency, because in the perspective of frequencist, the probability is the steady of the data, **so without repeat experiment, there is no probability**!
So, in the frequency family, they use booststrap method to determine the estimate error, Suppose our original data set consists of N data points X = {x1, . . . , xN}. We can create a new data set XB by drawing N points at random from X, with replacement, so that some points in X may be replicated in XB, whereas other points in X may be absent from XB. This process can be repeated L times to generate L data sets each of size N and each obtained by sampling from the original data set X. The statistical accuracy of parameter estimates can then be evaluated by looking at the variability of predictions between the different bootstrap data sets.
The criticisms of frequency family is mainly focus on three points:
1. The proposition of some problems is not justified
2. It is not appropriate to judge whether the statistical method is good or bad(pvalue).

5. Decision theory

We have seen the probability theory can offer a wonderful framework to measure uncertainty, and in this section, I want to discuss how to use decision theory to shed light on our understanding of prediction.

5.1 Regression

Suppose we have a real valued random input vector $X \in R^p$, and $Y \in R$ a real valued random output variable, with joint distribution $\Pr(X, Y)$. We seek a function $f(x)$ for predicting Y given by X. This theory requires a loss function $L(Y, f(x)) = (Y - f(x))^2$ for penalizing errors in prediction, and by far the most common and convenient is squared error loss: $L(Y, f(x)) = (Y - f(x))^2$. This leads us to a criterion for choosing f,

$$L(f) = E(Y - f(X))^2 = \int (y - f(x))^2 p(dx, dy)$$

Condition by X, we can rewrite our loss function

$$L(f) = E_X E_{Y|X} (Y - f(X))^2$$

And we see it is sufficient equal to minimize Loss function pointwise

$$f(x) = argmin_c E_{Y|X}\left([Y-c]^2 | X = x\right)$$

The solution is

$$f(x) = E(Y|X = x)$$

the conditional expectation, also known as the regression function. Thus the best prediction of Y at any point X = x is the conditional mean, when best is measured by average squared error.

Example. K nearest-neighbor

Reading. *Curse of dimensionality*

The meaning of Curse of dimensionality is that if the dimension of data getting higher, the local structure of the space is getting change. When the dimension p increases, the notion of "nearest points" vanishes. Suppose we need to use a huge vector $X = (X_1, \ldots, X_p) \in [0,1]^p$ follows a uniform distribution on the hypercube. And suppose we have two random variable R and R' follow uniform distribution on the [0,1], we have,

$$E\left[\left\|X^{(i)} - X^{(j)}\right\|^2\right] = \sum_{k=1}^{p} E\left[\left(X_k^{(i)} - X_k^{(j)}\right)\right] = pE[(R - R')] = p/6$$

And the standard deviation of this square distance is

$$sdev\left\|X^{(i)} - X^{(j)}\right\|^2 = \sqrt{pval[(R - R')]} \approx 0.2\sqrt{p}$$

In particular, we observe that the typical square distance between two points sampled uniformly in $[0,1]^p$ grows linearly with p, while the scaled deviation sdev/expectation shrinks like $p^{-1/2}$.

Furthermore, assuming the underlying regression function f0 is Lipschitz continuous, the knearest-neighbors estimate with $k \cong n^{2/(2+p)}$ satisfies

$$E\left\|\hat{f} - f_0\right\|^2 < n^{-2/(2+p)}$$

See Chapter 6.3 of Gyorfi et al. (2002)

Are we happy about L2 loss? Can we exchange other loss? What happens if we replace the L2 loss function with the L1: E|Y - f(X)|? The solution in this case is the conditional median

$$\hat{f}(x) = median(Y|X = x)$$

Squared error is analytically convenient so it is the most popular.

5.2 Classification

In the classification, our target is simple, is try to make less mistake. We need to define a rule to class each value of x into a suitable class. This rule will separate space into different region, those region is called decision region. Every category have a decision region, the boundary is called decision boundary or decision surface. Notebly, every decision region is not required to be continues.

First, let's consider the situation of two class. Like the instance in cancer, if we class the people belongs to C1(normal) into C2(or reversely), Then we make a mistake, than the probability of this thing happen is

$$p\left(mistake\right) = p\left(x \in R_1, C_2\right) + p\left(x \in R_2, C_1\right) = \int_{R_1} p\left(x, C_2\right) dx + \int_{R_2} p(x, C_1) dx$$

We want to minimize this functions, so if $\mathrm{p}(x, C_1) > p(x, C_2)$,than we need to class x into class C1. According to product rule, $\mathrm{p}(x, C_k) = p(C_k|x)p(x)$, so we need to class x into maximum posterior probability.

We consider the loss of different misclassification is equal, but for some situation, we may have different penalty on different type of misclassification.

Eample. Cancer vs normal classification

Our loss function can be represented by a K ×K matrix L, where K = card(C). L will be zero on the diagonal and nonnegative elsewhere, where L(k, ℓ) is the price paid for classifying an observation belonging to class Ck as Cℓ

So the loss function can be rewrite as

$$L\left(R\right) = E_X \sum_{k=1}^{K} L\left[C_k\ , R\left(X\right)\right] Pr\left(C_k|X\right)$$

Again we see that the k-nearest neighbor classifier directly approximates this solution—a majority vote in a nearest neighborhood amounts to exactly this, except that conditional probability at a point is relaxed to conditional probability within a neighborhood of a point, and probabilities are estimated by training-sample proportions.

Reference

Gyorfi, L., Kohler, M., Krzyzak, A. & Walk, H. (2002), A Distribution-Free Theory of Nonparametric Regression, Springer.

Young, Alastair G . Introduction to High-dimensional Statistics[J]. International Statistical Review, 2015, 83(3):515-516.

Ruppert D. The Elements of Statistical Learning: Data Mining, Inference, and Prediction[M]. 2008.

Bishop C M. Pattern Recognition and Machine Learning (Information Science and Statistics)[M]. 2006.