7. Bayesian Prediction

7.1 Bayes' Theorem

Bayesian inference derives the posterior probability as a consequence of two antecedents: a prior probability and a "likelihood function" derived from a statistical model for the observed data. Bayesian inference computes the posterior probability according to Bayes' theorem:

$$P(H \mid E) = \frac{P(E \mid H) \cdot P(H)}{P(E)}$$

where

- H stands for any hypothesis whose probability may be affected by data (called evidence below). Often there are competing hypotheses, and the task is to determine which is the most probable.
- $P(H)$, the prior probability, is the estimate of the probability of the hypothesis H before the data E, the current evidence, is observed.
- E, the evidence, corresponds to new data that were not used in computing the prior probability.
- $P(H \mid E)$, the posterior probability, is the probability of H given E, i.e., after E is observed. This is what we want to know: the probability of a hypothesis given the observed evidence.
- $P(E \mid H)$ is the probability of observing E given H, and is called the likelihood. As a function of E with H fixed, it indicates the compatibility of the evidence with the given hypothesis. The likelihood function is a function of the evidence, E, while the posterior probability is a function of the hypothesis, H.
- $P(E)$ is sometimes termed the marginal likelihood or "model evidence". This factor is the

same for all possible hypotheses being considered (as is evident from the fact that the hypothesis H does not appear anywhere in the symbol, unlike for all the other factors), so this factor does not enter into determining the relative probabilities of different hypotheses.

For different values of H, only the factors $P(H)$ and $P(E \mid H)$, both in the numerator, affect the value of $P(H \mid E)$ –the posterior probability of a hypothesis is proportional to its prior probability (its inherent likeliness) and the newly acquired likelihood (its compatibility with the new observed evidence).

Bayes' rule can also be written as follows:

$$P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)}$$

$$= \frac{P(E \mid H)P(H)}{P(E \mid H)P(H) + P(E \mid \neg H)P(\neg H)}$$

$$= \frac{1}{1 + \left(\frac{1}{P(H)} - 1\right)\frac{P(E|\neg H)}{P(E|H)}}$$

because

$$P(E) = P(E \mid H)P(H) + P(E \mid \neg H)P(\neg H)$$

and

$$P(H) + P(\neg H) = 1$$

where $\neg H$ is "not H", the logical negation of H. One quick and easy way to remember the equation would be to use Rule of Multiplication:

$$P(E \cap H) = P(E \mid H)P(H) = P(H \mid E)P(E)$$

阅读后结合上课内容掌握贝叶斯推理的基本思想，了解贝叶斯准则将两个变量概率传递的优雅方式。 推荐此部分的外部参考资源于 B 站 https://www.bilibili.com/video/BV15r4y1v7KT?spm_id_from=333.337.search-card.all.click，建议跟着流程看一遍。

7.2 Distribution-version of Bayesian Inference

Suppose:

- x, a data point in general. This may in fact be a vector of values.
- θ, the parameter of the data point's distribution, i.e., $x \sim p(x \mid \theta)$. This may be a vector of parameters.

- α, the hyperparameter of the parameter distribution, i.e., $\theta \sim p(\theta \mid \alpha)$. This may be a vector of hyperparameters.
- $\mathbf{X}$ is the sample, a set of n observed data points, i.e., $x_1, \ldots, x_n$.
- $\tilde{x}$, a new data point whose distribution is to be predicted.

Inference:

- The prior distribution is the distribution of the parameter(s) before any data is observed, i.e. $p(\theta \mid \alpha)$. The prior distribution might not be easily determined; in such a case, one possibility may be to use the Jeffreys prior to obtain a prior distribution before updating it with newer observations.
- The sampling distribution is the distribution of the observed data conditional on its parameters, i.e. $p(\mathbf{X} \mid \theta)$. This is also termed the likelihood, especially when viewed as a function of the parameter(s), sometimes written $\mathrm{L}(\theta \mid \mathbf{X}) = p(\mathbf{X} \mid \theta)$.
- The marginal likelihood (sometimes also termed the evidence) is the distribution of the observed data marginalized over the parameter(s), i.e.

$$p(\mathbf{X} \mid \alpha) = \int p(\mathbf{X} \mid \theta) p(\theta \mid \alpha) d\theta$$

.

- The posterior distribution is the distribution of the parameter(s) after taking into account the observed data. This is determined by Bayes' rule, which forms the heart of Bayesian inference:

$$p(\theta \mid \mathbf{X}, \alpha) = \frac{p(\theta, \mathbf{X}, \alpha)}{p(\mathbf{X}, \alpha)} = \frac{p(\mathbf{X} \mid \theta, \alpha) p(\theta, \alpha)}{p(\mathbf{X} \mid \alpha) p(\alpha)} = \frac{p(\mathbf{X} \mid \theta, \alpha) p(\theta \mid \alpha)}{p(\mathbf{X} \mid \alpha)} \propto p(\mathbf{X} \mid \theta, \alpha) p(\theta \mid \alpha).$$

  This is expressed in words as "posterior is proportional to likelihood times prior", or sometimes as "posterior = likelihood times prior, over evidence".
- In practice, for almost all complex Bayesian models used in machine learning, the posterior distribution $p(\theta \mid \mathbf{X}, \alpha)$ is not obtained in a closed form distribution, mainly because the parameter space for θ can be very high, or the Bayesian model retains certain hierarchical structure formulated from the observations $\mathbf{X}$ and parameter θ. In such situations, we need to resort to approximation techniques.

Prediction:

- The posterior predictive distribution is the distribution of a new data point, marginalized over the posterior:

$$p(\tilde{x} \mid \mathbf{X}, \alpha) = \int p(\tilde{x} \mid \theta) p(\theta \mid \mathbf{X}, \alpha) d\theta$$

- The prior predictive distribution is the distribution of a new data point, marginalized over the prior:

$$p(\tilde{x} \mid \alpha) = \int p(\tilde{x} \mid \theta) p(\theta \mid \alpha) d\theta$$

Bayesian theory calls for the use of the posterior predictive distribution to do predictive inference, i.e., to predict the distribution of a new, unobserved data point. That is, instead of a fixed point as a prediction, a distribution over possible points is returned. Only this way is the entire posterior distribution of the parameter(s) used. By comparison, prediction in frequentist statistics often involves finding an optimum point estimate of the parameter(s)—e.g., by maximum likelihood or maximum a posteriori estimation (MAP)—and then plugging this estimate into the formula for the distribution of a data point. This has the disadvantage that it does not account for any uncertainty in the value of the parameter, and hence will underestimate the variance of the predictive distribution.

In some instances, frequentist statistics can work around this problem. For example, confidence intervals and prediction intervals in frequentist statistics when constructed from a normal distribution with unknown mean and variance are constructed using a Student's t-distribution. This correctly estimates the variance, due to the facts that (1) the average of normally distributed random variables is also normally distributed, and (2) the predictive distribution of a normally distributed data point with unknown mean and variance, using conjugate or uninformative priors, has a Student's t-distribution. In Bayesian statistics, however, the posterior predictive distribution can always be determined exactly—or at least to an arbitrary level of precision when numerical methods are used.

Both types of predictive distributions have the form of a compound probability distribution (as does the marginal likelihood). In fact, if the prior distribution is a conjugate prior, such that the prior and posterior distributions come from the same family, it can be seen that both prior and posterior predictive distributions also come from the same family of compound distributions. The only difference is that the posterior predictive distribution uses the updated values of the hyperparameters (applying the Bayesian update rules given in the conjugate prior article), while the prior predictive distribution uses the values of the hyperparameters that appear in the prior distribution.

阅读后结合上课内容掌握贝叶斯在连续分布问题上的实现形式，对其中分布参数（θ 即是一般意义上的模型参数）的后验概率推导公式作深刻理解。

7.3 Frequentist

Frequentist inference is a type of statistical inference based in frequentist probability, which treats "probability" in equivalent terms to "frequency" and draws conclusions from sample-data by means of emphasising the frequency or proportion of findings in the data. Frequentist-inference underlies frequentist statistics, in which the well-established methodologies of statistical hypothesis testing and confidence intervals are founded.

The history of frequentist statistics is more recent than its prevailing philosophical rival, Bayesian statistics. Frequentist statistics were largely developed in the early 20th century and have recently

developed to become the dominant paradigm in inferential statistics, while Bayesian statistics were invented in the 19th century. Despite this dominance, there is no agreement as to whether frequentism is better than Bayesian statistics, with a vocal minority of professionals studying statistical inference decrying frequentist inference for being internally-inconsistent. For the purposes of this article, frequentist methodology will be discussed as summarily as possible but it is worth noting that this subject remains controversial even into the modern day.

The primary formulation of frequentism stems from the presumption that statistics could be perceived to have been a probabilistic frequency. This view was primarily developed by Ronald Fisher and the team of Jerzy Neyman and Egon Pearson. Ronald Fisher's contributed to frequentist statistics by developing the frequentist concept of "significance testing", which is the study of the significance of a measure of a statistic when compared to the hypothesis. Neyman-Pearson extended Fisher's ideas to multiple hypotheses by conjecturing that the ratio of probabilities of hypotheses when maximizing the difference between the two hypotheses leads to a maximization of exceeding a given p-value, and also provides the basis of type I and type II errors.

For statistical inference, the relevant statistic about which we want to make inferences is $y \in Y$, where the random vector Y is a function of an unknown parameter, θ. The parameter θ is further partitioned into (ψ, λ), where ψ is the parameter of interest, and λ is the nuisance parameter. For concreteness, ψ in one area might be the population mean, μ, and the nuisance parameter λ would then be the standard deviation of the population mean, σ.

Thus, statistical inference is concerned with the expectation of random vector Y, namely $E(Y) = E(Y;\theta) = \int y f_Y(y;\theta) dy$.

To construct areas of uncertainty in frequentist inference, a pivot is used which defines the area around ψ that can be used to provide an interval to estimate uncertainty. The pivot is a probability such that for a pivot, p, which is a function, that $p(t,\psi)$ is strictly increasing in ψ, where $t \in T$ is a random vector. This allows that, for some $0 < c < 1$, we can define $P\{p(T,\psi) \leq p_c^*\}$, which is the probability that the pivot function is less than some well-defined value. This implies $P\{\psi \leq q(T,c)\} = 1-c$, where $q(t,c)$ is a $1-c$ upper limit for ψ. Note that $1-c$ is a range of outcomes that define a one-sided limit for ψ, and that $1-2c$ is a two-sided limit for ψ, when we want to estimate a range of outcomes where ψ may occur. This rigorously defines the confidence interval, which is the range of outcomes about which we can make statistical inferences.

阅读后结合上课内容了解频率流派的基本思想与历史。

7.4 Testing for Statistical Independence

In this case, an "observation" consists of the values of two outcomes and the null hypothesis is that the occurrence of these outcomes is statistically independent. Each observation is allocated to

one cell of a two-dimensional array of cells (called a contingency table) according to the values of the two outcomes. If there are r rows and c columns in the table, the "theoretical frequency" for a cell, given the hypothesis of independence, is

$$E_{i,j} = N p_{i\cdot} p_{\cdot j},$$

, where N is the total sample size (the sum of all cells in the table), and

$$p_{i\cdot} = \frac{O_{i\cdot}}{N} = \sum_{j=1}^{c} \frac{O_{i,j}}{N},$$

, is the fraction of observations of type i ignoring the column attribute (fraction of row totals), and

$$p_{\cdot j} = \frac{O_{\cdot j}}{N} = \sum_{i=1}^{r} \frac{O_{i,j}}{N}$$

is the fraction of observations of type j ignoring the row attribute (fraction of column totals). The term "frequencies" refers to absolute numbers rather than already normalised values.

The value of the test-statistic is

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

$$= N \sum_{i,j} p_{i\cdot} p_{\cdot j} \left(\frac{(O_{i,j}/N) - p_{i\cdot} p_{\cdot j}}{p_{i\cdot} p_{\cdot j}} \right)^2$$

Note that χ^2 is 0 if and only if $O_{i,j} = E_{i,j} \forall i, j$, i.e. only if the expected and true number of observations are equal in all cells.

Fitting the model of "independence" reduces the number of degrees of freedom by $p = r + c - 1$. The number of degrees of freedom is equal to the number of cells rc, minus the reduction in degrees of freedom, p, which reduces to $(r-1)(c-1)$.

For the test of independence, also known as the test of homogeneity, a chi-squared probability of less than or equal to 0.05 (or the chi-squared statistic being at or larger than the 0.05 critical point) is commonly interpreted by applied workers as justification for rejecting the null hypothesis that the row variable is independent of the column variable. The alternative hypothesis corresponds to the variables having an association or relationship where the structure of this relationship is not specified.

拓展知识，阅读后简要理解何为统计独立性，以及研究其的意义。万事万物即变量，强行附会统计无关的变量只是徒劳。