과목명	데이터마이닝
담당교수	황두성 교수님
과제명	Homework 2
학과	소프트웨어학과
학번	32151671
이름	박민혁
제출일자	2021-05-18

1. Frequentism vs Bayesianism

Frequentism

- Frequency probability is one of probability analysis. This defines the probability of an event from the data obtained in many attempts as the limit of its relative frequency. In principle, probabilities can be found by repeatable objective processes. Like gambling, probabilities must be estimated without experiment. The need and development of these explanations of degree probabilities has been motivated by the problems and paradoxes of classical interpretation of probabilities, the previously dominant perspective. In classical interpretations, probability was defined as the principle of indifference or insufficient reason, depending on the natural symmetry of the problem. For example, the probability of a dice game arises from the natural symmetry of the cube, which is a cube. These classical interpretations are encountered in a natural unsymmetric statistical problem leading to inference.

Bayesianism

- Bayes' probability theory is a probability theory that interprets probability as 'an amount of knowledge or a degree of belief.' It is a different interpretation from considering probability as frequency of occurrence or physical properties of any system. There are two points of view in the Bayes probability: from an objective point of view, the Bayes statistics law can be proven reasonably universally and can be described as an extension of logic. Meanwhile, from the perspective of subjective probability theory, the state of knowledge can be measured to a degree of personal belief. Many modern machine learning methods are built on objective Bayes principles. Bayes probability theory is one of the most popular concepts of probability and is widely applied to psychology, sociology, and economics theory. To evaluate the probability of any hypothesis, we first reveal the prior probability and change the new probability value by the new relevant data. The Bayes theorem provides the necessary procedures and standards for the calculation of these probabilities.

2. What is the relationship between machine learning and frequentism?

Frequentism methods assume that the observed data were sampled from some distributions. We call this data distribution likelihood. $P(Data\,|\theta)$, Where θ is treated as a constant and the goal is to find θ to maximize likelihood. For example, logistic regression assumes that the data are sampled from a Bernoulli distribution, and linear regression assumes that the data are samples from a Gaussian distribution.

3. What is the relationship between machine learning and Bayesianism?

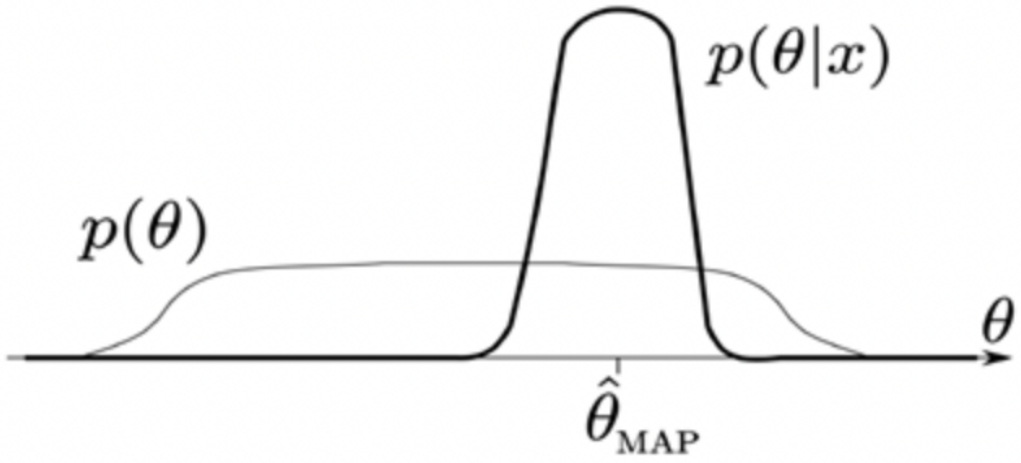
When machine learning is classified into large categories, it can be represented largely by regression and classification. The most basic linear regression is to infer the relationship between

independent and dependent variables. $y = \theta_0 + \theta_1 x$

The goal of regression is to minimize the difference between the estimates $y`$ and y created through these expressions. Machine learning is learned 'gradually' through algorithms such as Gradient Decent to find parameters. But let's change our gaze a little bit and think that the θ_0 and θ_1 we want to estimate do not have one specific value, but rather have a distribution. That way, we can express the process of machine learning finding parameters using the Bayes theorem.

$$P(model|data) = \frac{P(data|model)P(model)}{P(data)}$$

In other words, we know a prior called P(model) which is a machine learning process by obtaining posterior(P(model | data) when a new data is observed and gradually finding the distribution of P(model), or parameters, while using it as a prior for the next learning.



$$P(wi|x) = \frac{P(x|wi)P(wi)}{P(x)}$$

4. Test some code to compare frequentist and Bayesian approaches?

Photon Flux Measurements

First we will use Python to generate some toy data to demonstrate the two approaches to the problem. We will draw 50 samples F_i with a mean of 1000 (in arbitrary units) and a (known) error e_i:

```
>>> np.random.seed(2)

>>> e = np.random.normal
```

>>> F = np.random.normal

In this toy example we already know the true flux F, but the question is this: given our measurements and errors, what is our best point estimate of the true flux? Let's look at a frequentist and a Bayesian approach to solving this.

Frequentist Approach to Flux Measurement

We will start with the classical frequentist maximum likelihood approach. Given a single observation $D_i = (F_i, e_i)$, we can compute the probability distribution of the measurement given the true flux F given our assumption of Gaussian errors:

$$P(D_i \mid F) = (2\pi e^2{}_i)^{-1/2} \exp\left(\frac{-(F_i - F)^2}{2e^i{}_i}\right)$$

This should be read "the probability of D_i given F equals ...".

You should recognize this as a normal distribution with mean F and standard deviation e_i. We construct the *likelihood* by computing the product of the probabilities for each data point:

$$L(D \mid F) = \prod_{i=1}^{N} P(D_i \mid F)$$

Here $D = \{D_i\}$ represents the entire set of measurements. For reasons of both analytic simplicity and numerical accuracy, it is often more convenient to instead consider the log-likelihood; combining the previous two equations gives

$$logL(D \mid F) = -\frac{1}{2}\sum_{i=1}^{N}\left[\log(2\pi e^2{}_i) + \frac{(F_i - F)^2}{e^2{}_i}\right]$$

We would like to determine the value of F which maximizes the likelihood. For this simple problem, the maximization can be computed analytically (e.g. by setting $d\log L/dF|_{\hat{F}} = 0$),

which results in the following point estimate of F:

$$\hat{F} = \frac{\sum w_i F_i}{\sum w_i} \; ; \; w_i = 1 \,/\, e^2{}_i$$

The result is a simple weighted mean of the observed values. Notice that in the case of equal errors e_i, the weights cancel and $\hat{F}$ is simply the mean of the observed data.

We can go further and ask what the uncertainty of our esti- mate is. One way this can be accomplished in the frequentist approach is to construct a Gaussian approximation to the peak likelihood; in this simple case the fit can be solved analytically to give:

$$\sigma_{\hat{F}} = (\sum_{i=1}^{N} w_i)^{-1/2}$$

This result can be evaluated this in Python as follows:

```
>>> w = 1. / e ** 2
>>> F_hat = np.sum(w * F) / np.sum(w)
>>> sigma_F = w.sum() ** -0.5
```

For our particular data, the result is $\hat{F}$ = 999 ± 4.

Bayesian Approach to Flux Measurement

The Bayesian approach, as you might expect, begins and ends with probabilities. The fundamental result of interest is our knowledge of the parameters in question, codified by the probability *P*(*F*|*D*). To compute this result, we next apply Bayes' theorem, a fundamental law of probability:

$$P(F \mid D) = \frac{P(D \mid F)\, P(F)}{P(D)}$$

Though Bayes' theorem is where Bayesians get their name, it is important to note that it is not this theorem itself that is controversial, but the Bayesian *interpretation of probability* implied by the term *P*(*F*|*D*). While the above formulation makes sense given the Bayesian view of probability, the setup is fundamentally contrary to the frequentist philosophy, which says that probabilities have no meaning for fixed model parameters like *F*. In the Bayesian conception of probability, however, this poses no problem.

Let's take a look at each of the terms in this expression:

- *P*(*F*|*D*): The posterior, which is the probability of the model parameters given the data.
- *P*(*D*|*F*): The likelihood, which is proportional to the L (*D*|*F*) used in the frequentist approach.
- *P*(*F*): The model prior, which encodes what we knew about the model before considering the data *D*.
- *P*(*D*): The model evidence, which in practice amounts to simply a normalization term.

If we set the prior $P(F) \propto 1$ (a *flat prior*), we find $P(F|D) \propto L(D|F)$.

That is, with a flat prior on F, the Bayesian posterior is maximized at precisely the same value as the frequentist result! So despite the philosophical differences, we see that the Bayesian and frequentist point estimates are equivalent for this simple problem.

You might notice that we glossed over one important piece here: the prior, $P(F)$, which we assumed to be flat.[3] The prior allows inclusion of other information into the computation, which becomes very useful in cases where multiple measure- ment strategies are being combined to constrain a single model (as is the case in, e.g. cosmological parameter estimation). The necessity to specify a prior, however, is one of the more controversial pieces of Bayesian analysis.

A frequentist will point out that the prior is problematic when no true prior information is available. Though it might seem straightforward to use an uninformative prior like the flat prior mentioned above, there are some surprising subtleties involved.[4] It turns out that in many situations, a truly uninformative prior cannot exist! Frequentists point out that the subjective choice of a prior which necessarily biases the result should have no place in scientific data analysis.

A Bayesian would counter that frequentism doesn't solve this problem, but simply skirts the question. Frequentism can often be viewed as simply a special case of the Bayesian approach for some (implicit) choice of the prior: a Bayesian would say that it's better to make this implicit choice explicit, even if the choice might include some subjectivity. Further- more, as we will see below, the question frequentism answers is not always the question the researcher wants to ask.

5. What is the key impact of frequentism and Bayesianism in machine learning?

In the simple example above, the frequentist and Bayesian approaches give basically the same result. In light of this, arguments over the use of a prior and the philosophy of probability may seem frivolous. However, while it is easy to show that the two approaches are often equivalent for simple problems, it is also true that they can diverge greatly in other situations. In practice, this divergence most often makes itself most clear in two different ways:

1. The handling of nuisance parameters: i.e. param- eters which affect the final result, but are not other- wise of interest.

2. The different handling of uncertainty: for example, the subtle (and often overlooked) difference between frequentist confidence intervals and Bayesian credible regions.

Frequentism methods assume that the observed data were sampled from some distributions. We call this data distribution likelihood. $P(Data\,|\theta)$, Where θ is treated as a constant and the goal is to find θ to maximize likelihood. For example, logistic regression assumes that the data are sampled from a Bernoulli distribution, and linear regression assumes that the data are samples from a Gaussian distribution.

In other words, we know a prior called P(model) which is a machine learning process by obtaining posterior(P(model | data) when a new data is observed and gradually finding the distribution of P(model), or parameters, while using it as a prior for the next learning.