Contents

1. What We Talk About When We Talk About Bayesian Method

1.1 Bayesian vs. Frequentist, 1st Round

In general, what is “Bayesian” is “not frequentist”. As we have learned in class, the main disagreement between of Bayesian and frequentist, at least in statistics scenario, lies in that if the term “probability distribution of (true) model” (note as $p(f)$) makes sense. So, in a word, any method, for any purpose, in which $p(f)$ is used, it can be called a Bayesian method. Otherwise, it’s a frequentist method.

However, only the idea “$p(f)$ is reasonable” can’t help us to solve even toy problems. So many textbooks or online materials will also give us certain methods using $p(f)$ to solve all kinds of problems, which may confuse a beginner - “why these ‘Bayesian methods’ look so different”? And that the main motivation for me to write this note.

One more thing that should be clarified is that method using prior knowledge doesn’t have to be a Bayesian method. Frequentists don’t use prior distribution of model doesn’t mean they don’t use prior knowledge – they just use it in other not so evident ways. For example, in ridge linear regression we use regularization term $\lambda||\vec{w}||^2$, which is based on the prior knowledge that simpler model is more robust (i.e. Occam’s Razor. See lecture note 4 for more information).

If our model f is decided by a few parameters $\vec{\theta}$, $p(f)$ can be also written as $p(\vec{\theta})$ (for example, a univariate Gaussian distribution is decided by its expectation μ and variance σ^2). To keep it general, we would use $p(f)$ in most cases.

1.2 Topics

Naturally, we can see there are two main topics for Bayesian method:

1. “What Are We”: how to get $p(f)$;
2. “Where Are We Going”: how to use $p(f)$ to solve our problems.

In general, you can get your $p(f)$ by any kind of approach (for example?). After all, for Bayesians probability is just “degrees of belief”, isn’t it? In statistics/machine learning scenarios, we care more about “how to learn $p(f)$ from data”, or more specifically, “how to use data D to update prior distribution $p(f)$ and get posterior distribution $p(f|D)$”.

2. Updating $p(f)$ to $p(f|D)$

2.1 Bayesian Method and Bayes’ Rule

We said that any method using $p(f)$ can be called a Bayesian method, which doesn’t mention about Bayes’ rule at all. So, why we see Bayes’ rule in almost every specific approach solving problems? The answer lies in another sentence above: “how to use data D to update prior distribution $p(f)$ and get posterior distribution $p(f|D)$”.

We can interpret (continuous) Bayes' rule $p(f|D) = \frac{p(D|f)p(f)}{\int p(D|f)p(f)df}$ in two ways:

1. ("Belief Reweight") If we rewrite it as:

$$p(f|D) = \frac{p(D|f)}{\int p(D|f)p(f)df} \cdot p(f),$$

we would see that Bayes' rule updates $p(f)$ by product it with $\frac{p(D|f)}{\int p(D|f)p(f)df}$, which can be learned from data D. More precisely, it can be seen as "re-weighting the prior".

Noting $Z = \int p(D|f)p(f)df$ For a particular model f_0, by Bayes' rule:

$$p(f_0|D) = \frac{1}{Z} p(D|f_0)p(f_0)$$

Which means we will "amplify" f_0's weight if it is has a large likelihood $p(D|f_0)$, i.e. it is more likely to generate our observation data D.

2. ("Probability Inversion") If we rewrite it in another way:

$$p(f|D) = \frac{p(f)}{\int p(D|f)p(f)df} \cdot p(D|f),$$

we would see that we are turning likelihood $p(D|f)$ into posterior $p(f|D)$.

(See more in lecture note 1 about understanding Bayes' rule.)

The first form shows how Bayes' rule builds a bridge from prior $p(f)$ to posterior $p(f|D)$ by reweighting the prior, and that's why we see it so often. Sometimes, however, we may "skip" to use Bayes' rule. Recall the definition of conditional probability:

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

If we can estimate the joint distribution $p(f,D)$ directly, and then get $p(f|D)$ is almost a trivial work (theoretically, at least). Lecture note 6 gives us a such example, in which $\boldsymbol{y}$ and $\boldsymbol{w}$ are both Gaussian random vectors, so we can get their joint distribution $p\left(\begin{bmatrix}\boldsymbol{y}\\ \boldsymbol{w}\end{bmatrix}\middle| X\right)$ and then $p(w|D)$. Here we benefit from the good properties of Gaussian distribution, and we are still free to follow the standard Bayesian pipeline "likelihood-prior-posterior", maybe for fun.

2.2 Parametric and Nonparametric Method

As is shown in 1.1, If our model f is decided by a few parameters $\vec{\theta}$, $p(f)$ can be also written as $p(\vec{\theta})$. For example, if we study the distribution of a random variable X following normal distribution $\mathcal{N}(\mu,\sigma^2)$, then $p(f) = p(\vec{\theta}) = p(\mu,\sigma)$. Approaches based on this "parameterization" is called parametric approaches.

In some other cases, such as Gaussian process, we are faced with infinite, or extremely large number of parameters. Approaches used in these cases are called nonparametric approaches. Nonparametric approaches have a quite different pipeline with that of parametric approaches, so I decide to discuss them separately in another note.

2.3 Finding Likelihood $\mathrm{p}(D|f)$

As is shown in Bayes' rule, likelihood $p(D|f)$ is half of the main components of the bridge, and that's why we would like to study it.

For Bayesians, finding likelihood is to find the form instead of the value of parameters. One way to find likelihood is to study the properties of the experiment and make assumption. For example, we assume the likelihood of coin flipping to be $\mathrm{p}(\#_{\mathrm{head}} = \mathrm{x}|\mathrm{n}, \mathrm{p}_{head}) = \binom{n}{x} \mathrm{p}_{\mathrm{head}}^{\mathrm{x}}(1 - p_{\mathrm{head}})^{(n-x)}$. Here the form of likelihood is based on the physical properties of a coin and the setting of our experiment instead of data.

Another way is by transforming random variable with known probability distribution. In the Bayesian regression case, we know(assume) the weight $\vec{\mathrm{w}}$ and noise ϵ both follow Gaussian distribution: $\vec{w} \sim \mathcal{N}(\vec{\mu}, \Sigma)$ and $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$, with $\vec{w}, \epsilon$ and y having the relationship $\mathrm{y} = \mathrm{f}(\vec{\mathrm{w}}, \epsilon) = \mathrm{X}\vec{w} + \epsilon$. As Gaussian distribution in closed under linear transformation, y also follows Gaussian distribution: $\mathrm{p}(\mathrm{y}|\vec{w}) = \mathcal{N}(\mathrm{X}\mu, \mathrm{X}\,\Sigma\,\mathrm{X}^{\mathrm{T}} + \sigma^2 \mathrm{I})$.

Note here we benefit from the kindness of Gaussian distribution that it is closed under linear transformation, as well as many other operations. And that's one of the reasons we are likely to Gaussian distribution – it simplifies the computing greatly (another important reason would be about central limit theorem). However, (not surprisingly) there are some operations under which Gaussian distribution is not closed. In that case, approximation methods may help us. More of Gaussian distribution's properties are discussed in lecture note 5.

An interesting question may be: how does frequentists use likelihood? One answer is maximum likelihood estimator (MLE). MLE searches the its favorite parameter θ^*, which is "most likely" to generate the same data as ours, in entire parameter space Θ, i.e.:

$$\theta^* = \underset{\theta}{argmax}\ \ Likelihood(D|\theta)$$

In linear regression, the equation above would be:

$$\theta^* = \underset{\theta}{argmax}\ Likelihood(D|\theta) = \underset{w}{argmax}\ p(y|X, w)$$

If we assume $\mathbf{y_i}$s are independent given $\mathbf{x_i}$ and $\boldsymbol{w}$:

$$\theta^* = \underset{w}{argmax} \prod_{i=1}^{n} p(y_i|x_i, w) = \underset{w}{argmin} - \sum_{i=1}^{n} log\ p(y_i|x_i, w)$$

Another way of saying "finding the model that is most likely to generate the same data as ours" could be "finding the model generating data that are closest to ours by some metric". This idea lead to empirical risk minimization (ERM) and structural risk minimization (SRM). An example of ERM would be the ordinary least square regression:

$$\theta^* = \underset{\theta}{argmin} \sum_{i=1}^{n} \left(y_i - f(x_i, \theta)\right)^2$$

SRM takes the balance between the “likelihood”, or loss, and our prior knowledge about the model (e.g. natural models are likely to be simple). An example of SRM would be ridge regression:

$$\theta^* = \underset{\theta}{argmin} \sum_{i=1}^{n} \left(y_i - f(x_i, \theta)\right)^2 + \lambda \|\theta\|^2$$

Here we can actually tell the difference in “appearance” of frequentist and Bayesian method. The former is more likely to be a “search”, or optimization work, while the later struggles more with building probability models and computation issues.

2.4 Learning Posterior $\mathrm{p(f|D)}$

With likelihood $\mathrm{p(D|f)}$ and prior $\mathrm{p(f)}$, getting posterior $\mathrm{p(f|D)}$ may seems a trivial work – simply apply Bayes' rule and we’re done (see coin flipping case in lecture 2)!

Theoretical that’s true, but practically it may be a “simple but impossible” job. For example, I won’t worry about forgetting to bring an umbrella in rainy days if I can change the weather, but I just can’t. In many cases $\boldsymbol{p(D|f)p(f)}$ is quite hard to compute, not to mention the integral $\int \boldsymbol{p(D|f)p(f)df}$. And that’s why we need approximation method. Lecture note 8 introduces Laplace approximation.

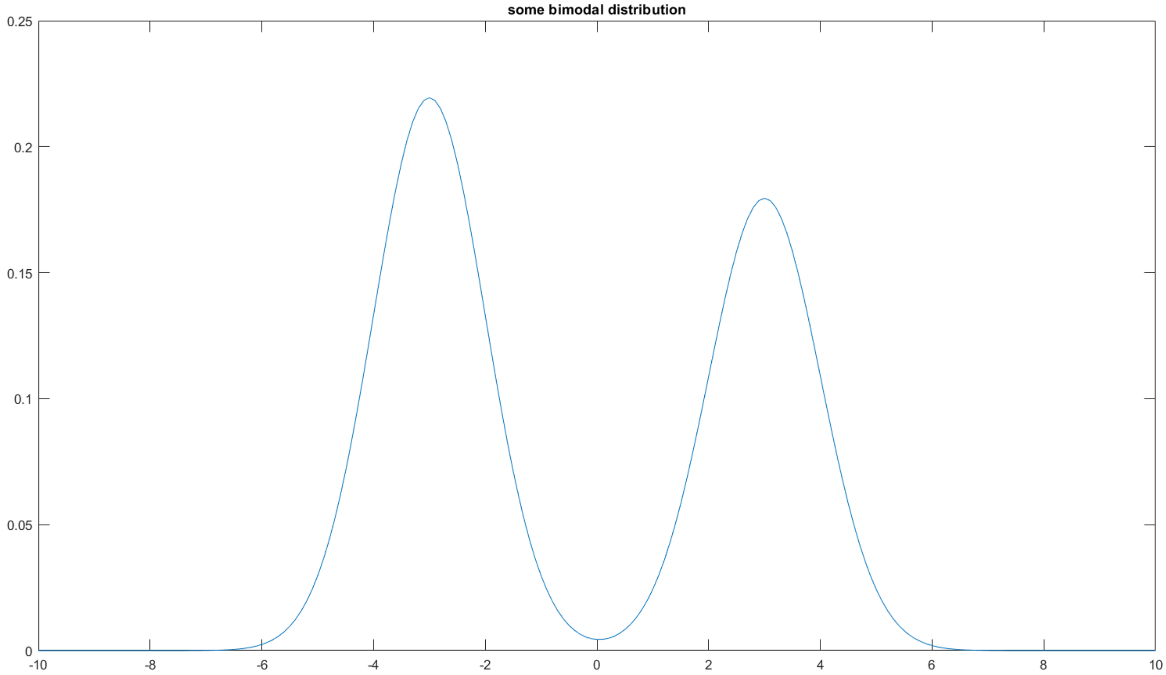
3. Using $\boldsymbol{p(f)}$: Parameter Estimation, Inference and Decision

In principle, we’re free to use $p(f)$ in any way we like to solve our problem. However, there’s a “True Bayesian” (or “typical Bayesian”) approach by “taking average”, which makes fully use of $p(f)$.

For example, in the parameter point estimation scenario, we do can use MAP estimator:

$$\theta^* = \underset{\theta}{argmax}\, p(\theta)$$

Sounds reasonable, right? But what if our $p(\theta)$ is bimodal and actually looks like:



In this case, the MAP estimator will be $\theta^* = -3$. However, we totally ignore the slightly lower peak at $\theta = 3$! So a better choice may be a weighted summation:

$$\theta^* = p(\theta = -3) * (-3) + p(\theta = 3) * 3$$

Basing on this idea we get the "true Bayes" estimator:

$$\theta^* = \int_\Theta \theta p(\theta) d\theta$$

In a word, if you are sure that there is a "best" estimator with absolute advantage over others, then "taking the best" strategy may work well and it's often easier to compute than "true Bayes" (no integral!). However, if there's not such estimator, "true Bayes" may give you a better result.

(By the way, this idea is somehow similar to that of ensemble learning, where we train a branch of classifiers and "averaging" them instead of picking out the best one.)

3.1 Parameter Estimation: Bayesian vs. Frequentist, 2nd Round

3.1.1 Point Estimation

We can use the coin flipping example again to show how Bayesian and Frequentist does parameter point estimation. This is not a question for frequentists, because once the estimator $\hat{\theta} = \frac{x}{n}$ is chosen, the only thing left is applying it on the data. Bayesians, however, are still free to choose their favorite point estimation criteria. For example, some may use maximum-a-posteriori (MAP) $\hat{\theta} = \underset{\theta}{argmax}\, p(\theta)$, while some other may use expectation $\hat{\theta} = E[\theta] = \int \theta p(\theta)\, d\theta$.

Does this show Bayesian method is better? Not really. Indeed, frequentists can do few things once a specific estimator is chosen, but they are free to choose the estimator! In fact there are many ways to evaluate the quality of an estimator, such as unbiasedness and consistency.

A related work is model selection, which is discussed in lecture 7. Actually the procedure of (Bayesian) model selecting is very similar to that of point estimation, so we don't discuss it here.

3.1.2 Interval Estimation

As is discussed in lecture 2, finding confidence interval by Bayesian method is almost trivial once we have $\mathrm{p(f)}$: for example, if we want to find a 95% confidence interval, we need only find $\mathrm{a, b} \in \mathrm{R}$ such that $\int_{\mathrm{a}}^{\mathrm{b}} \mathrm{p(f)df} = 95\%$. On the contrary, frequentists need to construct pivot variables and estimate them, which is often not so convenient as Bayesian method.

3.2 Inference: Bayesian vs. Frequentist, 3rd Round

For frequentists, once we learned the model f^* (i.e. learned θ^*), the only thing can be done is letting $\mathrm{y}_* = \mathrm{f}^*(x_*) = f(x_*, \theta^*)$. But for Bayesians. There are two ways:

1. Same as frequentists, first do a point estimation getting θ^*, and then compute $\mathrm{y}_* = f(x_*, \theta^*)$
2. "Model averaging": first compute probability distribution of y_*: $p(Y_*) = \int p(y_*|x_*, \theta)p(\theta|D)\, d\theta$, then do a "point estimation" on y_*, e.g. $\mathrm{y}_* = \mathrm{E}[Y_*]$.

In general, if you think single θ^* cannot "represent" $p(\theta)$ (e.g. $p(\theta)$ is bimodal and there are two very possible θ_1 and θ_2), then the second method would be a better choice. In model selection scenario, the second method would be $p(y_*|x_*, D) = \sum_i p(y_*|x_*, D, M_i)p(M_i, D)$, which is discussed in lecture 7.

3.3 Decision: Bayesian vs. Frequentist, Final Round

(to be added)

3.4 One More Thing: "Where Do We Come From"?

(About generative and discriminative model)

(to be added)