Nonparametric Bayes DRP Notes

Chai Harsha

Day 1 - 2/6/25

Definition. Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed (i.i.d.) random variables with common CDF F. The **empirical distribution** function is defined as

$$\hat{F}^n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i \leq x}.$$

Theorem (Law of Large Numbers)**.**

$$\lim_{n\to\infty} \hat{F}^n(x) \to F(x)$$

with probability 1.

Theorem (Central Limit Theorem)**.**

$$\lim_{n\to\infty} \sqrt{n}\hat{F}^n(x) \to N(F(x), F(x)(1-F(x)))$$

with probability 1.

Theorem (Glivenko-Cantelli Theorem)**.**

$$\lim_{n\to\infty} \hat{F}^n \to F$$

uniformly with probability 1.

$$P(\sup_x |\hat{F}^n(x) - F(x)| \to 0) = 1.$$

What is Probability?

Bayesian probability is a measure of the plausibility of an event given incomplete knowledge. Frequentist probability is a measure of the frequency of an event in a large number of trials. Both approaches can be applied to statistics.

Statistics

One truth μ, along with random data.

- Frequentists exclusively base their conclusions on repeated sampling.
- What if you can't smaple the data repeatedly? What is the probability that a team wins the Super Bowl in a given year?
- Bayesian argument - the level of belief in an event.
- In statistics, we have our observations $X_1, X_2, \ldots, X_n$ which are fixed, and we repeatedly update μ.
- To summarize, frequentists view the data is random and the truth is fixed, Bayesians fix the data while the truth is random.

Our framework is as follows:

$$\{X_i\}_{i=1}^n \sim p(\theta) = p(x|\theta)$$

where our prior distribution is $p(\theta)$ and our likelihood function is $p(x|\theta)$. The posterior distribution is

$$p(\theta|x) \propto \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)\,d\theta}.$$

If θ is a function, what is $p(\theta)$? If you can compute it, how do you compute $\int p(x|\theta)p(\theta)\,d\theta$?

Overview

- **Theory**

  1. Exchangeability - Our data is drawn from a conditional distribution, so we are really assuming that it is conditionally independent. $\{X_i\}_{i=1}^n$ are technically dependent! Di Finetti Theorem - Conditionally iid $\iff$ exchangeability.

  2. Frequentist guarantees - If we take the limit $n \to \infty$, we want to approach the truth. We can't know everything, so we need to know how close we are to the truth, even if the proof of this is finnicky.

- **Computation**

  1. Conjugacy - We can get around the integral $\int p(x|\theta)p(\theta)\,d\theta$ by choosing a prior that is conjugate to the likelihood function, which will save us from having to compute the integral analytically.

  2. MCMC - Markov Chain Monte Carlo - We can sample from the posterior distribution using MCMC methods.

Day 2 - 2/13/25

We study single-parameter models. There are four models which we will consider: binomial, normal, Poisson, and exponential.

1. **Binomial**

   We aim to estimate the population proportion from a sequence of Bernoulli trials (each data $y_1, \ldots, y_n \in \{0, 1\}$). Order does not matter (i.e. the data is **exchangeable**), so the model is defined by

   $$p(y|\theta) = \text{Bin}(y|n, \theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$$

   where θ is the probability of success, n is the number of trials, and y is the number of successes ($y \leq n$).

   Example (Probability of Female Birth)**.** We define θ to be the proportion of female births. Hence, $1 - \theta$ is the proportion of male births.Let y be the number of female births among n recorded births.

   We need a prior distribution for θ. For our purposes, $p(\theta) \sim \text{Unif}([0, 1])$.

From this, through Bayes' Law and removing constant terms w.r.t. the parameter, we obtain the posterior distribution

$$p(\theta|y) \propto \theta^y(1-\theta)^{n-y}.$$

However, in the case of a binomial distribution with uniform prior, we may explicitly calculate $p(y)$.

Once we have calculated the posterior, in order to make predictions under the above conditions, we have

$$\begin{aligned}\mathbb{P}(y_{n+1} = 1|y) &= \int_0^1 \mathbb{P}(y_{n+1} = 1|\theta, y)p(\theta|y)\,d\theta \\ &= \int_0^1 \theta \cdot p(\theta|y)\,d\theta \\ &= \mathbb{E}(\theta|y)\end{aligned}$$

The posterior incorporates information from the data, so it will be less variable than the prior. We formalize as the Tower Property:

$$\mathbb{E}(\theta) = \mathbb{E}(\mathbb{E}(\theta|y))$$

and

$$\mathrm{Var}(\theta) = \mathbb{E}(\mathrm{Var}(\theta|y)) + \mathrm{Var}(\mathbb{E}(\theta|y)).$$

How might we interpret the prior distribution? How might we select it?

- Population interpretation - the prior is a population of possible parameter values, from which the current was selected.
- State of knowledge interpretation - the prior distribution represents our knowledge about the parameter. A greater variance means that we know more about the underlying distribution.

A prior distribution that is of the same form as the posterior is called **conjugate**.

Day 3 - 2/20/25

The key will always be

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}.$$

For today, we will be using

$$p(h|D) = \frac{p(D|h)p(h)}{p(D)}.$$

The prior represents our information about our posterior. In many cases, the prior is uniform, which means

$$p(h|D) \propto p(D|h).$$

The *probability* of an event is

$$\int_{x-\delta}^{x+\delta} f(y)\,dy,$$

while the *likelihood* is just $f(x)$.

The *maximum likelihood estimator* (MLE) is, given $(X_n)_{i=1}^N \sim p(\cdot|\theta)$, is the parameter

$$\begin{aligned}\hat{\theta} &= \text{argmax}_\theta \prod_{i=1}^N p(X_i|\theta)\\ &= \text{argmax}_\theta \lim_{\delta\to 0} \prod_{i=1}^N \frac{\mathbb{P}_\theta(Y \in X_i \pm \delta)}{\delta}\end{aligned}$$

where $Y \sim p(\cdot, \theta)$. The steps are:

-1. Determine the model $p(\cdot, \theta)$. We are picking a class of function, for which θ is a parameter.

0. Generate $(X_i)_{i=1}^N$ from our model.

1. For each $\theta \in \mathbb{R}$, compute the likelihood of seeing $(X_i)_{i=1}^N$ using

$$\mathcal{L}(X, \theta) = \prod_{i=1}^N p(X_i|\theta).$$

2. Choose the θ that maximizes $\mathcal{L}$.

The *maximal a posteriori* (MAP) estimator is the same, but we maximize the posterior distribution instead of the likelihood function:

$$\hat{\theta}_{MAP} = \text{argmax}_\theta \prod_{i=1}^N p(\theta|D).$$

As we increase the amount of data we have to ∞, the MAP estimator converges to the MLE. Intuitively, the posterior is proportional to the likelihood, and with more data, the likelihood term dominates the prior.
Example. Let N_1 be the number of heads, and N be the total number of tosses. Let a, b be hyperparameters. Then

$$\begin{aligned}\hat{\theta}_{MAP} &= \text{argmax}_{\theta\in[0,1]} \frac{\theta^{N_1+a-1}(1-\theta)^{N-N_1+b-1}}{p(D)}\\ &= \text{argmax}_{\theta\in[0,1]} (N_1 + a - 1)\log(\theta) + (N - N_1 + b - 1)\log(1-\theta)\end{aligned}$$

We set

$$\frac{\partial g(\theta)}{\partial \theta} = 0.$$

$$\begin{aligned}0 &= \frac{N_1 + a - 1}{\theta} - \frac{N - N_1 + b - 1}{1-\theta}\\ &= (1-\theta)(N_1 + a - 1) - \theta(N - N_1 + b - 1)\\ \theta(N + a + b - 2) &= N_1 + a - 1\\ \hat{\theta}_{MAP} &= \frac{N_1 + a - 1}{N + a + b - 2}.\end{aligned}$$

We may also derive the MLE:

$$\begin{aligned}\hat{\theta}_{MLE} &= \text{argmax}_{\theta\in[0,1]}\theta^{N_1}(1-\theta)^{N-N_1} \\ &= \text{argmax}_{\theta\in[0,1]}(N_1)\log(\theta) + (N-N_1)\log(1-\theta)\end{aligned}$$

We set

$$\frac{\partial\mathcal{L}(\theta)}{\partial\theta} = 0.$$

We get

$$\begin{aligned}0 &= \frac{N_1}{\theta} - \frac{N-N_1}{1-\theta} \\ &= (1-\theta)N_1 - \theta(N-N_1) \\ \theta(N-N_1) &= (1-\theta)N_1 \\ \theta(N) &= N_1 \\ \hat{\theta}_{MLE} &= \frac{N_1}{N}.\end{aligned}$$