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

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

Day 4 - 2/27/25

Dirichlet Multinomial Model

Goal: generalize the Beta Binomial model Recall the Beta-Binomial model:

$$\begin{aligned}\theta &\sim \text{Beta}(\alpha,\beta) \\ X_i &\sim^{i.i.d.} \text{Binomial}(\theta)\,\forall\, i \in \{1:N\}\end{aligned}$$

Generative models, by the traditional definition, are a model for how the data is generated. For Beta-Binomial, we go top-down. We generate a θ drawn from our prior, and then generate X_i from the likelihood.

We can talk about the joint distribution,

$$p\left(\theta,(X_n)_{i=0}^{N}\right) = p_{\text{Beta}}(\theta)\prod_{i=1}^{N}\theta^{\mathbb{1}_{x_i=1}}(1-\theta)^{\mathbb{1}_{x_o=0}}.$$

We find the conditional distribution given the data from this.

For the Dirichlet-Multinomial model, we have, for $\alpha = (\alpha_1,\ldots,\alpha_k) \in \mathbb{R}_+^k$, we have

$$\begin{aligned}\theta = (\theta)_{i=1}^{k} &\sim \text{Dirichlet}(\alpha) \\ X_i &\sim^{i.i.d.} \text{Cat}(\theta)\,\forall\, i \in \{1:N\}\end{aligned}$$

where

$$\text{Dirichlet}(\alpha) \propto P(\theta)\alpha\prod_{c=1}^{k}\theta_c^{\alpha_c-1}$$

and

$$\text{Cat}(\theta) = \begin{cases} x_i = c \quad \text{with prob. } \theta_c \text{ where } c \in \{1 : K\} \end{cases}$$

We demand $\sum_{c=1}^{k} \theta_c = 1$. We call the

$$S_k = \left\{ \theta \in \mathbb{R}_+^k : \sum_{c=1}^{k} \theta_c = 1 \right\}$$

the k-dimensional simplex. Every point on the simplex is a probability distribution.

Then, the Dirichlet distribution is a *distribution over distributions*, since it is a distribution over the simplex.

Example. The Dirichlet distribution, for $k = 2$, is

$$\begin{aligned} p_{\text{Dir}(\alpha,\beta)} &\propto \theta_1^{\alpha-1}\theta_2^{\beta-1} \\ &= \theta_1^{\alpha-1}(1-\theta_2)^{\beta-1} \\ &\propto p_{\text{Beta}(\alpha,\beta)}(\theta_1) \end{aligned}$$

Then, altogether, the posterior is

$$\begin{aligned} p(\theta|D) &\propto p(\theta)p(D|\theta) \\ &\propto \prod_{c=1}^{k} \theta_c^{\alpha_c-1} \prod_{i=1}^{N} p(X_i|\theta) \\ &= \prod_{c=1}^{k} \theta_c^{\alpha_c-1} \prod_{i=1}^{N} \theta_c^{\mathbb{1}_{x_i=c}} \qquad \text{over all } c \\ &= \prod_{c=1}^{k} \theta_c^{\alpha_c + \sum_{i=1}^{N} \mathbb{1}_{x_i=c} - 1} \\ &= \text{Dirichlet}\left((\alpha_c + \text{num of } c\text{'s})_{c=1}^{k} \right) \\ &= \text{Dirichlet}\left(\alpha + \sum_{i=1}^{N} \mathbb{1}_{x_i=c} \right) \end{aligned}$$

Now, we seek to find

$$\hat{\theta}_{MAP} = \text{argmax}_\theta p(\theta|D)$$

$$\boxed{\textit{Exercise}}$$

Naive Bayes Classifier

We are trying to predict the labels y given data $\mathbf{x}$.

Example (Spam filtering)**.** Let $\mathcal{X}$ be the symbols you can type. Let $\mathbf{x} \in \mathbb{R}^d$ where d is the length of the email. Let $y \in \{0, 1\}$ where 1 indicates spam.

Goal: Given $D = \{(\mathbf{x}^i, y^i)\}_{i=1}^{N}$, we want to predict y^{N+1} given $\mathbf{x}^{N+1}$. In statistical terms, this is

$$p(y^{N+1}|D, \mathbf{x}^{N+1})$$

We invoke Naive Bayes. Let

$$\begin{aligned}
\pi &\sim \text{Beta}(\alpha, \beta) \\
y &\sim \text{Bernoulli}(\pi) \\
\mathbf{x} &\sim p(\mathbf{x}|y) \\
&\approx \prod_{i=1}^{d} p(x_i|y, \theta) && \text{(where } \theta \text{ is a hyperparameter)} \\
\text{where } x_i &\sim \text{Cat}(\theta|y) \\
\theta &\sim \text{Dirichlet}(\eta)
\end{aligned}$$

Day 5 - 3/6/25

Naive Bayes Classifier (*Cont.*)

Naive Bayes is a generative model. We would like to know how the data is generated.

Again, we have $D = \{\mathbf{x}^i, y^i\}_{i=1}^N$, where each $\mathbf{x} = (x_1, \ldots, x_D)$ and $y \in \{1, \ldots, C\}$.

We have that $y \sim \text{Dir}(\pi = (\pi_1, \ldots, \pi_C))$ where the Dirichlet distribution is $\text{Dir}(\pi \in S_d) \propto \prod_{i=1}^d \theta_i^{\pi_i - 1} \propto p(\theta)$ where S_d is the d-dimensional simplex and θ is a point on the simplex.

The parameter of our *frequentist* model is

$$\theta = \{\pi \in S_D, (\theta_{jc})\}$$

where

$$\begin{aligned}
y &\sim \pi \\
x_j | y = c &\sim p(x; \theta_{jc}) \, \forall \, j \in \{1 : D\}, \, \forall \, c \in \{1 : C\} && \text{(Condition on both feature and label)} \\
*p(x, y = c; \theta) &= \pi_c \prod_{j=1}^{D} p(x_j; \theta_{jc}) && \text{(Joint distribution)}
\end{aligned}$$

Given some data D as defined above, we can guess the parameters (assuming Bernoulli):

$$\begin{aligned}
\hat{\pi}_c &= \frac{\sum_{i=1}^N \mathbb{1}_{y^i = c}}{N} && \frac{\text{Num of class } c\text{'s in } D}{N} \\
\hat{\theta}_{jc} &= \frac{\sum_{i=1}^N \mathbb{1}_{y^i = c} \mathbb{1}_{x_j^i = 1}}{\sum_{i=1}^N \mathbb{1}_{y^i = c}} && \frac{\text{Num of yeses in the } j\text{th feature } D \text{ of class } c}{\text{Num of } j\text{th features in class } c} \text{ if } p(x_j; \theta_{jc}) = \text{Ber}(\theta_{jc})
\end{aligned}$$

Here, we assumed some god-given parameters, and we want to choose estimators that are asymptotically close to the true parameters.

For the *Bayesian* model, we have hyperparameter

$$H = \{\alpha \in S_C, \beta_{jc}^1, \beta_{jc}^2\}$$

where

$$\begin{aligned}
\pi &\sim \mathrm{Dir}(\alpha) \\
y &\sim \pi \\
\theta_{jc} &\sim \mathrm{Beta}(\beta^1_{jc}, \beta^2_{jc})\,\forall\, j \in \{1:D\},\,\forall\, c \in \{1:C\} \\
x_j|y = c &\sim \mathrm{Ber}(\theta_{jc}) \\
*p(x, y = c|\pi, \theta) &= \pi_c \prod_{j=1}^{D} p(x_j|\theta_{jc}) = \pi_c \prod_{j=1}^{D} \theta_{jc}^{\mathbb{1}_{x_j=1}}(1-\theta_{jc})^{\mathbb{1}_{x_j=0}}
\end{aligned}$$

Our MAP estimator is then

$$\begin{aligned}
&p(\pi|D) && \text{(posterior on } \pi) \\
&p(\theta_{jc}|D) && \text{(posterior on } \theta_{jc}) \\
p(y = c|\mathbf{x}, D) &\propto p(x|y, D)p(y|D)dx && \text{(posterior predictive)}
\end{aligned}$$

Day 6 - 3/13/25

Latent Variable Models

In these models, we have a discrete latent state,

$$z_i \in \{1, \dots, K\}.$$

We can use a discrete prior, say $p(z_i) = \mathrm{Cat}(\pi)$. For the likelihood, we can use

$$p(x_i|z_i = k) = p_k(x_i),$$

where p_k is the base distribution for some k. This is the mixture model, given as

$$p(x_i|\theta) = \sum_{k=1}^{K} \pi_k p_k(x_i|\theta).$$

Then, π can be thought of as a set of weights that normalize the sum $\sum_{k=1}^{K} p_k(x|\theta)$.

Now, **Gaussian mixture models** are models in which each $p_k(x_i|\theta) = \mathcal{N}(x_i|\mu_k, \Sigma_k)$.

We also have **Multinomial mixture models**, given by

$$p(x_i|z_i = k, \theta) = \prod_{j=1}^{D} \mathrm{Ber}(x_{ij}|\mu_{jk}) = \prod_{j=1}^{D} \mu_{jk}^{x_{ij}}(1-\mu_{jk})^{1-x_{ij}}.$$

Clustering

We would like to use mixture models to cluster data. We would like to find which cluster is "responsible" for each data point. That is,

$$r_{ik} = p(z_i = k|x_i, \theta) = \frac{p(z_i = k|\theta)p(x_i|z_i = k, \theta)}{\sum_{k=1}^{K} p(z_i = k|\theta)p(x_i|z_i = k, \theta)}.$$

Note: Finding the correct K is reserved for nonparametric Bayes, for now we say that it is God-given.

We solve this by the **Expectation-Maximization** algorithm. We start with some θ, and then we iterate:

1. **E-step:** We compute r_{ik} for all i, k.
2. **M-step:** We update θ by maximizing the likelihood given the r_{ik} values.

Intuitively, the E-step is computing which mean is the best for each datum, and the M-step is updating the means to be the best for the data.

Remark 0.1. As an aside, the dat ais generated as such:

$$\begin{aligned} \pi &\sim \text{Dir}(\alpha) \\ z_i &\sim \text{Cat}(\pi) \\ (\mu_k, \Sigma_k) &\sim \text{Conjugate of Normal} \\ x_i | z_i = k &\sim \mathcal{N}(\mu_k, \Sigma_k) \end{aligned}$$

Markov Chain Monte Carloc (MCMC)

How do we actually sample from the posterior? We use MCMC methods.

Let $\mathcal{S}$ denote the state space. Let $\{X_i\}_{i=1}^n$ be a Markov chain with transition kernel P. That is, it satisfies

$$\mathbb{P}(X_{n+1} = j | X_n, \dots, X_1) = \mathbb{P}(X_{n+1} = j | X_n) = P_{X_n, j}.$$

Theorem. *If $\{X_i\}_{i=1}^n$ is aperiodic and irreducible, then*

$$\lim_{n \to \infty} \mathbb{P}(X \in A \subseteq \mathcal{S}) = \pi(A)$$

where $A = \{x_i : i = 1, \dots, m\}$ and $\pi(A) = \sum_{i=1}^m \pi(i)$.

The goal of MCMC is, given some π, can we construct a Markov Chain $\{X_i\}_{i=1}^n$ such that π is the stationary distribution of the chain?

The answer is yes, under minimal technical conditions.