1 Recap: Posterior Probability

\[ P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)} \]

- \( P(\theta|x) \): Posterior.
- \( P(x|\theta) \): Likelihood.
- \( P(\theta) \): Prior.
- \( P(x) = \int P(x|\theta)P(\theta)d\theta \): Marginal likelihood.

Most of Bayesian inference is about computing the posterior, or computing statistics on the posterior.

1.1 Predictive probability

Once we observed a bunch of data, we may want to compute probability of new point

\[ P(x_{new}|x) = \int P(x_{new}|\theta)P(\theta|x)d\theta \]

2 How do you specify/select prior

Assume prior was Uniform.

\[ P(\theta) = C \quad \text{where } C = \text{constant} \]

When you assume a Uniform prior, things get simplified. The posterior,

\[ P(\theta|x) = \frac{P(x|\theta)C}{\int P(x|\theta)C d\theta} \]

The posterior becomes proportional to the likelihood

\[ P(\theta|x) \propto P(x|\theta) \]
2.1 Problems with Uniform priors

2.1.1 Problem 1

Consider a Gaussian distribution

\[ P(x|\mu, \sigma^2) \]

We want to put a prior on \( \mu \). We use uniform prior \( P(\mu) = C \). Here we have a problem with the choice of this prior, what is that?

In order to get proper probability distribution,

\[ \int_{-\infty}^{\infty} P(\mu)d\mu = 1 \]

But in the above case of uniform prior, we are going to get

\[ \int_{-\infty}^{\infty} P(\mu)d\mu = +\infty \]

which is not a proper probability distribution. This is one possible reason why it is not reasonable to apply these kinds of prior.

*Note: If you have an improper prior, it does not necessarily mean the posterior is improper. In fact, in many cases, the posterior is proper. From the point of view of doing Bayesian analysis, we can deal with the improper prior as long as the posterior is proper distribution.*

2.1.2 Problem 2: Issue of change of variables

If we change the variables, i.e if we re-parameterize the distribution, the re-parameterized distribution will be non-uniform if it is non-linear change of variables. Consider Bernoulli distribution, which is parameterized in terms of \( \theta \)(fraction of heads)

\[ \text{odds ratio}(r) = \frac{\theta}{1-\theta} \]

\[ P(r) = P(\theta)|\frac{d\theta}{dr}| \]

Here \( \frac{d\theta}{dr} \) is non-uniform and and \( P(\theta) \) is uniform, so \( P(r) \) is not uniform. So, we could go from uniform to non-uniform prior and it is dependent of how we choose to parameterize the distribution, which is not very good. It does not make sense if it is based on the parameterizations.

2.2 Quick Detour: Decision theory

In decision theory, we have a few different elements

- \( P(\theta) \): Parameter which is a fixed unknown quantity unlike in Bayesian analysis.
- \( P(x|\theta) \): Likelihood.
- \( \delta(x) \): decision (procedure) Eg: MLE.
- \( l(\theta, \delta(x)) \): loss function which is the loss between \( \theta \) and decision

*Eg: Assume the decision be MLE i.e \( \delta(x) = \hat{\theta}_{MLE} \) and the loss function : \( (\theta - \hat{\theta}_{MLE})^2 \). This loss function tells how close is the estimated \( \theta \) from the true \( \theta \)?*
2.2.1 Frequentist risk

Expected loss when we run experiments over and over again.

\[ r(\theta, \delta(x)) = \int l(\theta, \delta(x))P(x|\theta)dx \]

This can be explained as: drawing x from distribution, computing decision (eg: MLE), computing loss, repeat computing the loss over and over again and finally, averaging the overall loss. *Note:* There is no prior over here, there is just the likelihood.

2.2.2 Posterior risk

From a Bayesian context: Instead of averaging over various data sets, average over \( \theta \)

\[ \rho(\theta, \delta(x)) = \int l(\theta, \delta(x))P(\theta|x)dx \]

We fix the data (no need to care about other data sets). We look at the loss under different choice of parameters. i.e we are averaging over different choice of parameters and looking at the average loss. This is what is studied in Bayesian decision theory, where we will look at average loss under the posterior.

2.2.3 Bayes action(\( \delta^*(x) \))

The decision that minimizes the posterior risk under a particular choice of loss function. Say, the loss function:

\[ l(\theta, \delta(x))) = (\theta - \delta(x))^2 \]

We want to figure out the decision that minimizes the posterior risk under squared loss. The posterior risk

\[ \rho(\theta, \delta(x)) = \int (\theta - \delta(x))^2P(\theta|x)d\theta \]

\[ \frac{d\rho}{d\delta(x)} = 2\delta(x) - 2\int \theta P(\theta|x)d\theta \]

To find the Bayes action, set \( \frac{d\rho}{d\delta(x)} = 0 \). We get,

\[ \delta^*(x) = \int \theta P(\theta|x)d\theta \]

which is nothing but a posterior mean \( E[\theta|x] \).

3 Conjugate priors

**Definition 1** (Conjugate prior). Given a likelihood, the *conjugate prior* is the prior distribution such that the prior and posterior are in the same family of distributions.
3.1 Bernoulli distribution

\[ P(x|\theta) = \theta^x (1 - \theta)^{1-x} \]

We want to put prior on \( \theta \).

The conjugate prior for Bernoulli is \( \text{Beta} \) distribution. Beta is probability distribution on two parameters \( \alpha \) and \( \beta \):

\[ P(\theta|\alpha,\beta) = \frac{\theta^{\alpha-1} (1 - \theta)^{\beta-1}}{B(\alpha,\beta)} \]

where \( B(\alpha,\beta) = \int_0^1 \theta^{\alpha-1} (1 - \theta)^{\beta-1} d\theta \) is \( \text{Beta} \) function \( (B(\alpha,\beta)) \). So,

\[ P(\theta|\alpha,\beta) \propto \theta^{\alpha-1} (1 - \theta)^{\beta-1} \]

\( P(\theta|\alpha,\beta) \) is proportional to \( \theta^{\alpha-1} (1 - \theta)^{\beta-1} \) (in terms of \( \theta \))

3.2 Why Beta is Conjugate prior for Bernoulli?

Now we show that, if the prior is a Beta distribution, the likelihood is bernoulli distribution then the posterior is also a beta distribution. (which is the definition if conjugate prior)

\[ P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)} \]

\[ P(\theta|x) = \frac{\theta^x (1 - \theta)^{1-x} \theta^{\alpha-1} (1 - \theta)^{\beta-1}}{P(x)B(\alpha,\beta)} \]

\[ P(\theta|x) \propto \theta^{\alpha+x-1} (1 - \theta)^{\beta+(1-x)-1} = \text{Beta}(\hat{\alpha}, \hat{\beta}) \]

where \( \hat{\alpha} = \alpha + x \) and \( \hat{\beta} = \beta + (1 - x) \). Assume that we flipped a coin. If we flipped heads \( (x=1) \), increment \( \alpha \) by 1 and if we flipped tails \( (1-x=1) \), increment \( \beta \) by 1. So, we are just incrementing either \( \alpha \) or \( \beta \) in the bernoulli trial. We can generalize this argument for sequence of coin flips. \( X=X_1, x_2, ..., x_n \). The likelihood

\[ P(X|\theta) = \prod_i P(x_i|\theta) \]

If we have Beta prior and above likelihood, the posterior is \( \text{Beta}(\hat{\alpha}, \hat{\beta}) \) where \( \hat{\alpha} = \alpha + \#\text{heads} \) and \( \hat{\beta} = \beta + \#\text{tails} \). So, in the sequence of coin flips, this is how we would update the parameters of the Beta distribution.

3.3 Gamma \( \Gamma \) function

Beta function

\[ B(\alpha,\beta) = \int_0^1 \theta^{\alpha-1} (1 - \theta)^{\beta-1} d\theta \]

Gamma function

\[ \Gamma(z) = \int_0^\infty \exp^{-t} t^{z-1} dt \]
We can write Beta in terms of the Γ function
\[ B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)} \]

Γ function as generalization of factorial function (recursive property of γ function):
\[ \Gamma(z) = (z - 1)! \]
\[ \Gamma(z + 1) = z\Gamma(z) \]

where z is a positive integer.

### 3.4 Mean of Beta distribution

Mean is the expected value of Beta distribution
\[
\int_0^1 \theta P(\theta | \alpha, \beta) d\theta = \int_0^1 \theta^{\alpha-1}(1 - \theta)^{\beta-1} d\theta
\]
\[
Mean = \frac{\text{Beta}(\alpha + 1, \beta)}{\text{Beta}(\alpha, \beta)} = \frac{\alpha}{\alpha + \beta} \text{ [by using recursive property]}
\]
which is the expected value of Beta distribution.

### 3.5 Posterior Mean in Bernoulli distribution

Posterior mean in a Beta Bernoulli
\[
\hat{\alpha} = \frac{\alpha + \#heads}{\alpha + \beta + n}
\]

If we have a series of coin flips, and we want to do Bayesian inference, this is what we can predict as our estimate for θ. We can write this as following:
\[
w(\frac{\alpha}{\alpha + \beta}) + (1 - w)\hat{\theta}_{MLE}
\]
where \(w = \frac{\alpha + \beta}{n + \beta + n} \in (0, 1)\). The above equation is a weighted combination between a prior mean and a MLE. As \(n\) goes to \(\infty\), \(w\) goes to zero. This says that, as we observe more data, the estimator for the posterior mean becomes increasingly dominated by MLE. In the limit, as \(n\) approaches to \(\infty\), the posterior mean approaches MLE. And MLE has a property, which is called statistical consistency, which says that as we get more and more data, the MLE approaches true θ. So the posterior mean is also a consistent estimator.

### 3.6 Binomial distribution

Binomial says: If we have a biased coin, and we flip it \(n\) times, the probability of getting \(m\) heads
\[
P(m | n, \theta) = \binom{n}{m} (\theta)^m (1 - \theta)^{n-m} \propto (\theta)^m (1 - \theta)^{n-m}
\]
So, we can use same prior (Beta) that we used in Bernoulli for Binomial as well. The posterior in Beta \(\hat{\alpha} = \alpha + m\) and \(\hat{\beta} = \beta + (n - m)\) which are the same posterior estimate we got in the Bernoulli.
### 3.7 Categorical / Discrete distribution

Instead of looking at binary outcome, we look at the one of k-distribution (e.g., rolling a weighted die with k-sides, and we have some probability of each outcome). \( \theta_1, \theta_2, ..., \theta_k \) where \( \sum_i \theta_i = 1 \)

\[
P(x|\theta) = \prod_i (\theta_i)^{x_i}
\]

So, we can generalize the Bernoulli to k different outcomes. \( P(x_i = k) = \theta_k \)

### 3.8 Multinomial distribution

Generalizing binomial distribution gives multinomial distribution,

\[
P(x_1, x_2, ..., x_k|\theta, n) = \frac{n!}{\prod_i x_i! \prod_i (\theta_i)^{x_i}}
\]

### 3.9 Priors of categorical and multinomial distribution

We generalize the Beta distribution to k different outcomes.

\[
P(\theta|\alpha) \propto \prod_i (\theta_i)^{\alpha_i-1}
\]

where \( \alpha \) is a k-dim vector.

The normalization constant is multinomial beta function:

\[
B(\alpha) = \frac{\prod_i \Gamma(\alpha_i)}{\Gamma(\sum_i \alpha_i)}
\]

### 3.10 Dirichlet distribution

Conjugate prior for multinomial and categorical distributions:

\[
P(\theta|\alpha) = \frac{\prod_i \Gamma(\alpha_i)}{\Gamma(\sum_i \alpha_i)} \prod_i (\theta_i)^{\alpha_i-1}
\]

We can view the Dirichlet distribution as a distribution over distributions. A sample from Dirichlet gives a group of \( \theta \)'s that sum to one. This is discrete distributions, a probability distribution over discrete objects.

Expectation of single \( \theta_i \) is:

\[
E[\theta_i|\alpha] = \int \theta_i P(\theta|\alpha) d\theta
\]

\[
E[\theta_i|\alpha] = \int \theta_i \frac{\prod_i \Gamma(\alpha_i)}{\prod_i \Gamma(\alpha_j)} \prod_i (\theta_i)^{\alpha_i-1} \prod_j (\theta_j)^{\alpha_j-1} d\theta
\]

\[
E[\theta_i|\alpha] = \frac{\Gamma(\sum_j \alpha_j)}{\prod_j \Gamma(\alpha_j)} \int \theta_i (\theta_i)^{\alpha_i-1} (\prod_j (\theta_j)^{\alpha_j-1}) d\theta
\]

\[
E[\theta_i|\alpha] = \frac{\Gamma(\sum_j \alpha_j)}{\prod_j \Gamma(\alpha_j)} \theta_i (\theta_i)^{\alpha_i-1} (\prod_j (\theta_j)^{\alpha_j-1}) - \Gamma(\sum_j \alpha_j)
\]
In the above equation, the integral part is multinomial beta function over $\hat{\alpha}$, where $\hat{\alpha}_j = \alpha_j$ and $\hat{\alpha}_i = \alpha_i + 1$ ($i \neq j$)

$$E[\theta_i | \alpha] = \frac{\Gamma(\sum_j \alpha_j) \prod_j \Gamma(\hat{\alpha}_j)}{\prod_j \Gamma(\alpha_i) \Gamma(\sum_j \hat{\alpha}_j)}$$ (4)

$$= \frac{\Gamma(\sum_j \alpha_j)}{\Gamma(\sum_j \alpha_j + 1)} \frac{\Gamma(\alpha_i + 1)}{\Gamma(\alpha_i)}$$ (5)

$$= \frac{\alpha_i}{\sum_j \alpha_j} \text{[by using recursive property: } \Gamma(z + 1) = z\Gamma(z)] \text{ (6)}$$

The other question we can ask: what is the posterior of the Dirichlet when the likelihood being multinomial or categorical:
If we have a multinomial likelihood, $P(x|\theta)$ where $x$ is number of successes of each of the $k$-trials, and then we have a prior $P(\theta)$, then the posterior is a Dirichlet prior with $(\alpha + x)$ i.e. $P(\theta | \alpha + x)$. 