50008 - Probability and Statistics - Lecture 10

Oliver Killane

09/03/22

Posterior

Lecture Recording

Lecture recording is available here

MLE Sensitivity

There are several shortcomings of **MLE**:

- **Sensitive to Sample Size**
  In a small sample, small fluctuations can change the **MLE** considerably.
- **Does not use any Prior Information**
  Only uses the given sample.
- **Returns a single value**
  Only returns the single and specific value $\hat{\theta}$, not a distribution $P(\theta|\underline{x})$ for some sample $\underline{x}$.

  Hence we cannot know how close other θ are, how strong our estimate is.
- **Cannot Assess**
  Can only assess using confidence intervals, however these are also dependent on the sample.

Bayes & Posterior

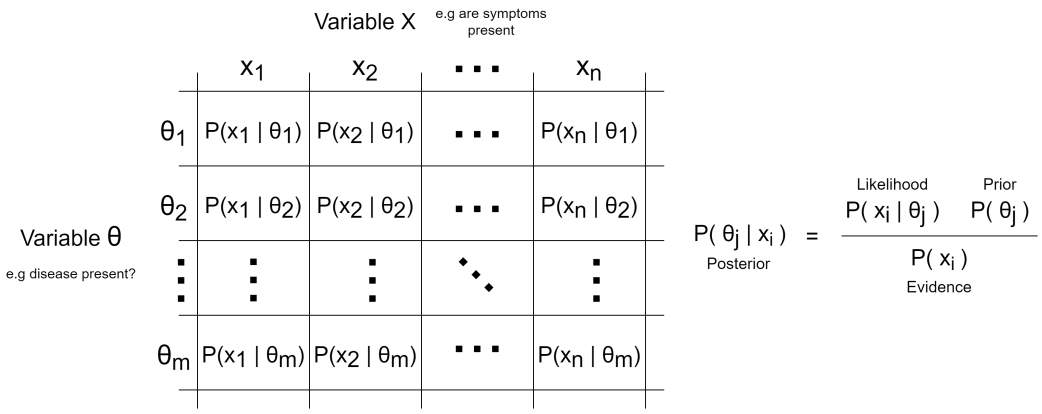
Definition: Baye's Theorem

Given two events A and B, where $P(B) \neq 0$:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

Note that we can use the law of total probability to re-express this without knowing $P(B)$:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B|A) \times P(A) + P(B|\overline{A})(1 - P(A))}$$



By law of total probability:

$$\text{Given } j \in [1,m].\ \sum_{i=1}^{n} P(x_i|\theta_j) = 1 \ \text{ and given } i \in [1,n] \sum j = 1^m P(\theta_j|x_i) = 1$$

When calculating the **MLE** using a sample $\underline{x}$ we calculated:

$$\hat{\theta}_{MLE} = arg\max_{\theta} L(\theta|\underline{x}) = arg\max_{\theta} \left[\prod_{i=1}^{n} P(x_i|\theta)\right]$$

(The θ most likely to give the sample $\underline{x}$)

We can apply this to the distributions X and θ to get a joint distribution:

$$P(\theta|X) = \frac{P(X|\theta) \times P(\theta)}{P(X)}$$

Where the **evidence** (X), acts as a normalizer (does not alter the shape of the distribution, just stretches/compresses it to normalize so that the distribution of $\theta|X$ has total probability 1)

$$\int_{-\infty}^{\infty} P(\theta|X)d\theta = 1$$

Hence we can say that the likelihood, and the posterior are directly proportional:

$$P(\theta|X) \propto P(X|\theta)P(\theta)$$

Maximum a Posteriori (MAP) Estimate

Definition: Maximum a Posteriori Estimate (MAP Estimate)

Given some prior information $(P(\theta))$ we can effectively get the **MLE**, but each probability is weighted by the prior information.

$$\begin{aligned}
\hat{\theta}_{MAP} &= arg\max_{\theta} \left[\prod_{i=1}^{n} P(\theta|X = x_i)\right] \\
&= arg\max_{\theta} \left[\prod_{i=1}^{n} \frac{P(X = x_i|\theta) \times P(\theta)}{P(X = x_i)}\right] \\
&= arg\max_{\theta} \left[\prod_{i=1}^{n} P(X = x_i|\theta) \times P(\theta)\right]
\end{aligned}$$

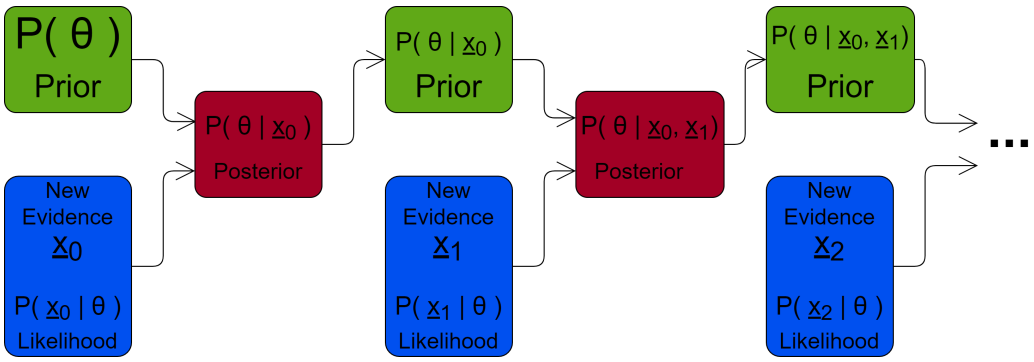
Using the uniform distribution as $P(\theta)$ yields the **MLE** as each $P(X = x_i|\theta)$ is equally weighted.

Conjugate Priors

Lecture Recording

Lecture recording is available here

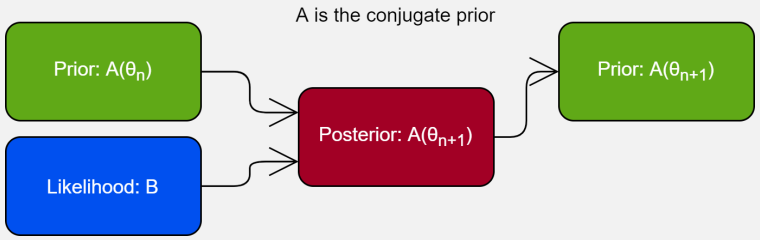
We can continually use the **MAP** to get new prior information, to use with new evidence to refine the **MAP**. This process of continually using the previous estimate and new evidence to refine the estimate is called **Baysian Inference**



$$\text{where } P(\theta|X) = \frac{P(X|\theta) \times P(\theta)}{P(X)} = \frac{P(X|\theta) \times P(\theta)}{\int_{-\infty}^{\infty} P(X|\theta)P(\theta)\ d\theta}$$

Definition: Conjugate Prior

When continually inferring new prior distributions, if the prior distribution is in the same family of distributions (i.e parameters can be different, but same distribution) as the posterior, then it is a **conjugate prior**.



Likelihood	Conjugate Prior
Bernoulli Binomial Geometric	Beta
Poisson Exponential	Gamma
Normal	Normal

Definition: Beta Prior Distribution

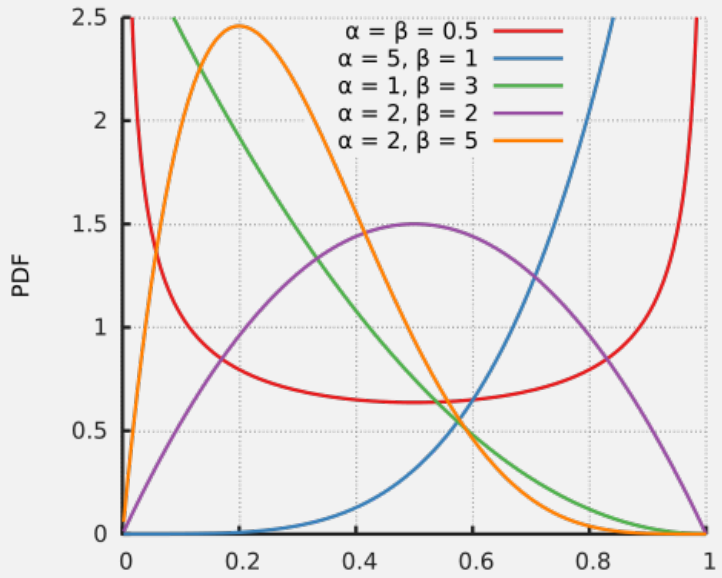
Where $\alpha, \beta > 0$ are **hyper-parameters** that determine the shape of the distribution, the parameter is θ:

$$Beta(\theta; \alpha, \beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}$$

Where the normalising value (ensures total integral sums to 1 so it is a valid **pdf**) is:

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

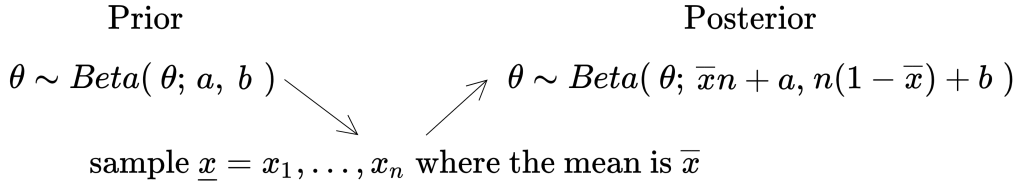
maximal value/θ_{MAP}	**mean/bayesian estimate θ_B**	**variance**
$argmax_\theta[Beta(\theta; \alpha, \beta)]$	$E[\theta]$	$E[\theta^2] - (E[\theta])^2$
$m_{\alpha,\beta} = \frac{\alpha - 1}{\alpha + \beta - 2}$	$\mu_{\alpha,\beta} = \frac{\alpha}{\alpha + \beta}$	$\sigma^2_{\alpha,\beta} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$



- When $\alpha = \beta$ it is symmetrical about 0.5
- higher values result in steeper/narrower distribution
- The MAP estimate pulls the estimate towards the prior.
- As $\alpha \to 1$ and $\beta \to 1$ $Beta(\theta; \alpha, \beta) \to U(0, 1)$ and $\hat{\theta}_{MAP} \to \hat{\theta}_{MLE}$.

Computing Terms

Bernoulli Distribution



Given some $x_i|\theta \sim Bernoulli(\theta)$ we choose the conjugate pair as $\theta \sim Beta(\theta; \alpha, \beta)$ where $\alpha > 1$ and $\beta > 1$.

We have a sample from the distribution: $\underline{x} = x_1, x_2, \ldots, x_n$

Step 1. Given $\theta \sim Beta(\theta; \alpha, \beta)$, the sample $\underline{x} = x_1, x_2, \ldots, x_n$ and sample mean $\overline{x}$ we need to calculate:

$$P(\theta|\underline{x}) = \frac{P(\underline{x}|\theta)P(\theta)}{P(\underline{x})} = \frac{P(\underline{x}|\theta)P(\theta)}{\int_{-\infty}^{\infty} P(\underline{x}|\theta)P(\theta)\ d\theta}$$

We know that the number of 1s in the sample is $\overline{x}n$.

Step 2. First we calculate $P(\underline{x}|\theta)P(\theta)$ using the bernoulli **PMF**:

$$\begin{aligned} P(\underline{x}|\theta) &= \prod_{i=1}^{n} P(x_i|\theta) \\ &= \theta^{\overline{x}n}(1-\theta)^{n-\overline{x}n} \\ &= \theta^{\overline{x}n}(1-\theta)^{n(1-\overline{x})} \end{aligned}$$

By the pdf of the **Beta** distribution:

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

Where B is the beta distribution normalization.

Hence we can multiply to get $P(\underline{x}|\theta)P(\theta)$:

$$\begin{aligned} P(\underline{x}|\theta)P(\theta) &= \theta^{\overline{x}n}(1-\theta)^{n(1-\overline{x})}\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} \\ &= \frac{\theta^{\overline{x}n+\alpha-1}(1-\theta)^{n(1-\overline{x})+\beta-1}}{B(\alpha,\beta)} \end{aligned}$$

Step 3. We derive $P(\theta|\underline{x})$:

$$P(\theta|\underline{x}) = \frac{P(X|\theta)P(\theta)}{P(\int_{-\infty}^{\infty} P(X|\theta)P(\theta))\ d\theta}$$

$$= \frac{\frac{\theta^{\overline{x}n+\alpha-1}(1-\theta)^{n(1-\overline{x})+\beta-1}}{B(\alpha,\beta)}}{\int_{-\infty}^{\infty} \frac{\theta^{\overline{x}n+\alpha-1}(1-\theta)^{n(1-\overline{x})+\beta-1}}{B(\alpha,\beta)}\ d\theta}$$

$$= \frac{\frac{\theta^{\overline{x}n+\alpha-1}(1-\theta)^{n(1-\overline{x})+\beta-1}}{B(\alpha,\beta)}}{\frac{1}{B(\alpha,\beta)}\int_{-\infty}^{\infty} \theta^{\overline{x}n+\alpha-1}(1-\theta)^{n(1-\overline{x})+\beta-1}\ d\theta}$$

$$= \frac{\theta^{\overline{x}n+\alpha-1}(1-\theta)^{n(1-\overline{x})+\beta-1}}{\int_{-\infty}^{\infty} \theta^{\overline{x}n+\alpha-1}(1-\theta)^{n(1-\overline{x})+\beta-1}\ d\theta}$$

$$= P(\theta) \quad \text{given} \quad \theta \sim Beta(\theta; \overline{x}n+\alpha, n(1-\overline{x})+\beta)$$

Hence we have the posterior distribution:

$$\theta \sim Beta(\theta; \overline{x}n+\alpha, n(1-\overline{x})+\beta)$$

New Bayesian Estimate

The new bayesian estimate is a **convex combination** of the **sample mean** $\overline{x}$ and the prior mean (prior bayesian estimate).

$$\begin{aligned}\hat{\theta}_B &= \frac{\overline{x}n+\alpha}{\overline{x}n+\alpha+n(1-\overline{x})+\beta} \\ &= \frac{\overline{x}n+\alpha}{\alpha+n+\beta} \\ &= \left(\underbrace{\overline{x}}_{\hat{\theta}_{MLE}} \times \frac{n}{n+\alpha+\beta}\right) + \left(\underbrace{\frac{\alpha}{\alpha+\beta}}_{\text{old } \hat{\theta}_B = \mu_{\alpha,\beta}} \times \frac{\alpha+\beta}{n+\alpha+\beta}\right)\end{aligned}$$

Lecture Recording

Lecture recording is available here

Normal Distribution - Single DataPoint Sample

Given some $x|\mu \sim N(\mu, \sigma^2)$ where σ^2 is known and μ is unknown. Using a sample of a single datapoint x.

Step 1. The likelihood can be found using the **Normal Distribution PDF**:

$$\begin{aligned} P(x|\mu) &= f(x|\mu) \\ &= \frac{1}{\sigma\sqrt{2\pi}} \times exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \text{ where } exp\{n\} = e^n \end{aligned}$$

Hence we now need to calculate the prior (the previous μ value that we will update with our estimate, using the sample):

$$\mu \sim N(\mu_0, \sigma_0^2)$$

Hence we can now calculate the **posterior distribution**.

Step 2. We calculate the **posterior distribution**

$$P(\mu|x) = f(\mu|x) = \frac{f(x|\mu)f(\mu)}{f(x)} = \frac{f(x|\mu)f(\mu)}{\int_{-\infty}^{\infty} f(x|\mu)f(\mu)\ d\mu}$$

$$\vdots$$

$$= (\text{some constant}) \times exp\left\{-\frac{\left(\mu - \frac{\mu_0\sigma^2 + x\sigma_0^2}{\sigma^2 + \sigma_0^2}\right)^2}{2 \times \frac{\sigma^2\sigma_0^2}{\sigma^2 + \sigma_0^2}}\right\}$$

We can express the new variance as:

$$\sigma_1^2 = \left(\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}\right)^{-1} \quad \text{and} \quad \mu_1 = \sigma_1^2\left(\frac{\mu_0}{\sigma_0^2} + \frac{x}{\sigma^2}\right)$$

With the new posterior density function as:

$$\mu|X \sim N(\mu_1, \sigma_1^2)$$

Normal Distribution - Sample Size n

We extend the previous proof for a sample $\underline{x} = x_1, \ldots, x_n$ and distribution $x_i|\mu \sim N(\mu, \sigma^2)$ where σ is known.

Step 1. We calculate the likelihood:

$$\begin{aligned}
P(\underline{x}|\mu) = f(\underline{x}|\mu) &= f(x_1|\mu)f(x_2|\mu)\dots f(x_n|\mu) \\
&= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\} \\
&= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \times \prod_{i=1}^{n} exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\} \\
&= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \times exp\left\{\sum_{i=1}^{n} -\frac{(x_i-\mu)^2}{2\sigma^2}\right\} \\
&= \frac{1}{\sigma^n(2\pi)^{n/2}} \times exp\left\{\sum_{i=1}^{n} -\frac{(x_i-\mu)^2}{2\sigma^2}\right\}
\end{aligned}$$

And then the prior probability which is distributed by $\mu \sim N(\mu_0, \sigma_0^2)$.

$$P(\mu) = f(\mu) = \frac{1}{\sigma_0\sqrt{2\pi}} exp\left\{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\}$$

Step 2. We can then calculate the posterior using **baye's theorem** .

$$\begin{aligned}
P(\mu|\underline{x}) &= \frac{1}{\sigma_0\sqrt{2\pi}} exp\left\{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\} \times \frac{1}{\sigma^n(2\pi)^{n/2}} \times expr\left\{\sum_{i=1}^{n} -\frac{(x_i-\mu)^2}{2\sigma^2}\right\} \\
&= \frac{1}{(2\pi)^{(n+1/2)}\sigma_0\sigma^n} exp\left\{\frac{-\mu^2+2\mu\mu_0-\mu_0^2}{2\sigma_0^2} - \sum_{i=1}^{n}\frac{x_i^2-2\mu x_i+\mu^2}{2\sigma^2}\right\} \\
&\vdots \\
&\propto exp\left\{-\frac{\left(\mu - \frac{\mu_0\sigma^2+\sum_{i=1}^{n}\sigma_0^2 x_i}{\sigma^2+n\sigma_0^2}\right)^2}{2\frac{\sigma_0^2\sigma^2}{\sigma^2+n\sigma_0^2}}\right\}
\end{aligned}$$

Hence we have:

$$\mu|\underline{x} \sim N(\mu_1, \sigma_1^2)$$

$$\sigma_1^2 = \frac{\sigma^2\sigma_0^2}{\sigma^2+n\sigma_0^2} = \left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)^{-1} \quad \text{and} \quad \mu_1 = \frac{\mu_0\sigma^2+\sum_{i=1}^{n}\sigma_0^2 x_i}{\sigma^2+n\sigma_0^2} = \sigma_1^2\left(\frac{\mu_0}{\sigma_0^2}+\sum_{i=1}^{n}\frac{x_i}{\sigma^2}\right)$$

Normal Distribution - Sufficient Statistic

Definition: Sufficient Statistic

A statistic is **sufficient** for a given model (our chosen distribution) and its associated parameter if no other statistic can be calculated from a sample that provides additional information in computing the value/estimate of the unknown parameter.

For a **normal distribution** the sufficient statistic is the sample mean:

$$T(\underline{x}) = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Hence we will use the sample mean in calculating our posterior distribution.

Step 1. We can directly calculate the posterior distribution using the likelihood and prior.

$$\begin{aligned}
P(\mu|\underline{x}) = f(\mu|\underline{x}) &= \frac{f(\mu)f(\underline{x}|\mu)}{\int_{-\infty}^{\infty} f(\mu)f(\underline{x}|\mu)\ d\mu} \\
&\propto \frac{f(\mu)f(T(\underline{x})|\mu)}{\int_{-\infty}^{\infty} f(\mu)f(\underline{x}|\mu)\ d\mu} \\
&\propto f(\mu)f(T(\underline{x})|\mu) \\
&= f(\mu)f(\overline{x}|\mu) \\
&= \frac{1}{\sigma_0\sqrt{2\pi}} exp\left\{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\} \times \frac{1}{\sqrt{2\pi\dfrac{\sigma^2}{n}}} exp\left\{-\frac{n(\overline{x}-\mu)^2}{2\sigma^2}\right\} \\
&\vdots \\
&\propto exp\left\{\frac{-\left(\mu - \dfrac{\mu_0\sigma^2/n + \overline{x}\sigma_0^2}{\sigma^2/n + \sigma_0^2}\right)^2}{2\dfrac{\sigma_0^2\sigma^2/n}{\sigma^2/n + \sigma_0^2}}\right\}
\end{aligned}$$

Hence we have the exponential part of the pdf for a normal distribution.

Step 2. We can now compute the posterior distribution.

$$\mu|\underline{x} \sim N(\mu_1, \sigma_1^2)$$

$$\sigma_1^2 = \frac{\sigma_0^2\sigma^2/n}{\sigma^2/n + \sigma_0^2} = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1} \quad \text{and} \quad \mu_1 = \frac{\mu_0\sigma^2/n + \overline{x}\sigma_0^2}{\sigma^2/n + \sigma_0^2} = \sigma_1^2\left(\frac{\mu_0}{\sigma_0^2} + \frac{\overline{x}n}{\sigma^2}\right)$$