3. Parametric Density Estimation

Dongwoo Kim

dongwookim.ac.kr

CSED515 - 2023 Spring

Table of Contents

Motivation

We have a coin, and a probability to get a head of the coin is μ.

- We flipped the coin 10 times and observed 7 heads and 3 tails.
- Q: what would be the most plausible value of μ given these observations?

Motivation

We have a coin, and a probability to get a head of the coin is μ.

- We flipped the coin 10 times and observed 7 heads and 3 tails.
- Q: what would be the most plausible value of μ given these observations?
  - A: $\mu = 0.7$

Motivation

We have a coin, and a probability to get a head of the coin is μ.

- We flipped the coin 10 times and observed 7 heads and 3 tails.
- Q: what would be the most plausible value of μ given these observations?
  - A: $\mu = 0.7$
- Where did you get this number? Are there any other answers?
- How can we formalize this process (observation $\rightarrow$ parameter) in a principled way?

Statistical Model: a Set of Probabilistic Models

One way to extract patterns from data is to find the most likely probability model generating observed data $\mathcal{D}$ among a set of probabilistic models (or statistical model):

- Supervised learning
  - Use samples of input x and output y, i.e., $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$
  - Learn a mapping from input to output under a probabilistic model of $p(y \mid x)$, e.g., a parameterized model $p(y \mid x, \theta)$
- Unsupervised learning (today; for simplicity)
  - Use samples of input x, i.e., $\mathcal{D} = \{x_i\}_{i=1}^n$
  - Learn an explanation using a probabilistic model of $p(x)$, e.g., a parameterized model $p(x \mid \theta)$

Application of Statistical Model

- Predicting the expectation $\mathbb{E}[X]$ or variance $\mathrm{Var}[X]$
  - c.f., considering (x, y) as a sample x', the learned joint distribution $p(x' = (x, y))$ allows us to predict $\mathbb{E}[Y \mid X]$
- Predicting the tail distribution $\inf\{a : p(X \geq a) \leq 0.1\}$, ... [1]
- Detecting outliers (a.k.a. ood; out-of-distribution) by checking likelihood $p(X = x_*)$
- ...

[1] https://en.wikipedia.org/wiki/Infimum_and_supremum

A Typical Setup for Statistical Model (1)

- Let $X_1, ..., X_n$ be n independent copies of X, i.e., X_i's are drawn from a single distribution independently (i.i.d.)
- The goal of statistics is to learn the distribution of X

A Typical Setup for Statistical Model (1)

- Let $X_1, \ldots, X_n$ be n independent copies of X, i.e., X_i's are drawn from a single distribution independently (i.i.d.)
- The goal of statistics is to learn the distribution of X
- e.g., survey on the number of siblings:

$$0, 2, 0, 1, 2, 3, 0, 1, \ldots$$

- We could make no assumption and try to learn the pmf:

x	0	1	2	3	4	5	≥ 6
$p(X = x)$	p_0	p_1	p_2	p_3	p_4	p_5	$p_{6^+} = \sum_{i \geq 6} p_i$

where we need to learn 7 parameters (count & normalize)

A Typical Setup for Statistical Model (2)

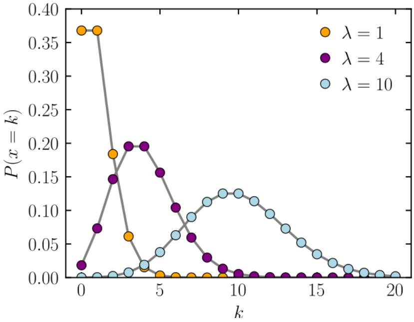


Figure: PMF of Poisson distribution

- Instead, we could assume that $X \sim \text{Poisson}(\lambda)$ with single parameter[2]

[2] `https://en.wikipedia.org/wiki/Poisson_distribution`

Density Estimation

- The density estimation is the problem of modeling a probability density function $p(x)$ given a finite number of data points, $\{x_i\}_{i=1}^n$ drawn from that density function
- Approaches to density estimation
  - Parametric estimation (this lecture) assumes a specific functional form for density model governed by a set of parameters, and finds the most likely parameters that explain the data.
  - Nonparametric estimation has no specific function form, and allows the form of the density to be determined entirely by the data, e.g., histogram[3], kernel density estimation.

[3] https://en.wikipedia.org/wiki/Histogram

Parameter Estimation (1)

Definition (Parametric statistical model)

Let the observed outcome of a statistical experiment be a sample $X_1, ..., X_n$ of n i.i.d. random variables in some measurable space Ω (usally $\Omega \subseteq \mathbb{R}$ and denote by p their common distribution. A statistical model associated to that statistical experiment is a pair

$$(\Omega, (p_\theta)_{\theta \in \Theta}) \ ,$$

where

- Ω is sample space
- $(p_\theta)_{\theta \in \Theta}$ is a family of probability measures on Ω, e.g., Bernoulli, Gaussian, ...
- $\Theta \subseteq \mathbb{R}^d$ is parameter set (for some $d \geq 1$)

Parameter Estimation (2)

- Usually, we will assume that the statistical model is well specified, i.e., $\exists \theta_* \in \Theta$ s.t. $p = p_{\theta_*}$
- This particular θ_* is called the true parameter, and is unknown
- The aim of the statistical experiment is to estimate θ_*, or check it's properties when they have a special meaning, e.g., $\theta > 1$? or $\theta \neq 1/2$?, ...
- But, the fundamental problem is finding $\hat{\theta} \approx \theta_*$, where the quality of approximation is often measured[5] by
  - Bias $(\mathbb{E}_D[\hat{\theta}] - \theta_*)$ and variance $\mathbb{E}_D[(\mathbb{E}_D[\hat{\theta}] - \hat{\theta})^2]$
  - Note that if $\Theta \subseteq \mathbb{R}$, risk $=$ bias$^2 +$ variance

[4] hat (^) indicates an estimated value in general.

[5] Note that $\hat{\theta}$ is a random variable.

Additional Slides: Bias-Variance Tradeoff

To measure the quality of estimator, we use Risk $=$ Bias$^2+$ Variance. Where does this come from?

From mean squared error of an estimator,

$$\begin{aligned}
&\mathbb{E}[(\hat{\theta}-\theta_*)^2] = \mathbb{E}[\hat{\theta}^2 - 2\theta_* \cdot \hat{\theta} + \theta_*^2] \\
&= \mathbb{E}[\hat{\theta}^2] - 2\theta\,\mathbb{E}[\hat{\theta}] + \theta_*^2 \\
&= \mathbb{E}[\hat{\theta}^2] - 2\theta\,\mathbb{E}[\hat{\theta}] + \theta_*^2 + \mathbb{E}^2[\hat{\theta}] - \mathbb{E}^2[\hat{\theta}] \\
&= \underbrace{(\mathbb{E}[\hat{\theta}] - \theta_*)^2}_{\text{Bias}^2} + \underbrace{\mathbb{E}[\hat{\theta}^2] - \mathbb{E}^2[\hat{\theta}]}_{\text{Var}(\hat{\theta})}
\end{aligned}$$

Table of Contents

Maximum Likelihood Estimation (MLE)

- The likelihood function $\mathcal{L}(\theta; \mathcal{D}) := p(\mathcal{D} \mid \theta) = p_\theta(\mathcal{D})$ expresses how probable the observation is for different values of parameter θ
- MLE finds the parameters $\hat{\theta}_{\mathsf{MLE}}$ maximizing the likelihood function, i.e.,

$$\hat{\theta}_{\mathsf{MLE}} := \arg\max_{\theta} \mathcal{L}(\theta; \mathcal{D}) \; .$$

Maximum log-Likelihood Estimation (MLE)

The log-likelihood $\ell(\theta; \mathcal{D}) := \log(\mathcal{L}(\theta; \mathcal{D}))$ is often used

- Since that log is monotonically increasing, we have

$$\hat{\theta}_{\text{MLE}} \;:=\; \arg\max_{\theta} \mathcal{L}(\theta; \mathcal{D}) \;=\; \arg\max_{\theta} \ell(\theta; \mathcal{D}) \; .$$

- Suppose each point of $\mathcal{D} = \{x_1, ..., x_n\}$ is drawn independently from $p(\cdot \mid \theta)$. Then, we have $p(\mathcal{D} \mid \theta) = \prod_{i=1}^{n} p(x_i \mid \theta)$ and thus

$$\ell(\theta; \mathcal{D}) = \sum_{i=1}^{n} \log(p(x_i \mid \theta)) \; .$$

An Example of MLE: Binomial distribution

Assume we have observed x heads out of n trials of a coin flip from $\text{Bin}(x|\mu, n)$ with unknown μ. Then, MLE solution maximizes the following loss function:

$$\begin{aligned} \ell_{\text{MLE}}(\mu) &= \log p(x|\mu) \\ &= \log \binom{n}{x} \mu^x (1-\mu)^{n-x} \\ &\propto x \log \mu + (n-x) \log(1-\mu) \end{aligned}$$

Then, it follows from solving $\frac{\partial \ell_{\text{MLE}}}{\partial \mu} = 0$ that

$$\hat{\mu}_{\text{MLE}} = \frac{x}{n} \, .$$

An Example of MLE: Gaussian (1)

Suppose that we wish to estimate μ from its noisy observation $x_i \underset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ for $i = 1, ..., n$

- Estimator 1: takes the first sample only, i.e., $\hat{\mu} = x_1$, then

$$\mathbb{E}[\hat{\mu}] = \mu \ , \quad \text{and} \quad \text{Var}(\hat{\mu}) = \sigma^2 \ ,$$

- Estimator 2: takes the average, i.e., $\overline{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$, then

$$\mathbb{E}[\overline{\mu}] = \mu \ , \quad \text{and} \quad \text{Var}(\overline{\mu}) = \frac{\sigma^2}{n} \ ,$$

Both estimators are unbiased, i.e., $\mathbb{E}[\hat{\mu}] = \mathbb{E}[\overline{\mu}] = \mu$, but

$$\text{Var}(\overline{\mu}) \leq \text{Var}(\hat{\mu}) \ ,$$

i.e., the risk of $\overline{\mu} = 0^2 + \sigma^2/n$ is smaller than the risk of $\hat{\mu} = 0^2 + \sigma^2$

An Example of MLE: Gaussian (2)

- It turns out that the empirical mean $\overline{\mu} = \mu_{\mathsf{MLE}}$.

- From now on, we will obtain the MLE solution $\hat{\theta}_{\mathsf{MLE}}$ s.t.

$$\hat{\theta}_{\mathsf{MLE}} = (\hat{\theta}_{\mathsf{MLE},1}, \hat{\theta}_{\mathsf{MLE},2}) \quad \approx \quad \theta = (\theta_1, \theta_2) = (\mu, \sigma^2)$$

- The parameterized density $p(x \mid \theta)$ is given by

$$p(x \mid \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) .$$

- The log-likelihood with $\theta = (\theta_1, \theta_2) = (\mu, \sigma^2)$ is given as

$$\ell(\theta) = \sum_{i=1}^{n} \log p(x_i \mid \theta) = \sum_{i=1}^{n} \left[-\frac{1}{2}\log(2\pi\theta_2) - \frac{1}{2\theta_2}(x_i - \theta_1)^2\right]$$

An Example of MLE: Gaussian (3)

We find stationary points by solving $\nabla_\theta \ell(\theta) = 0$:

$$\frac{1}{\theta_2}\sum_{i=1}^{n}(x_i - \theta_1) = 0 \; , \quad \text{and} \quad -\sum_{i=1}^{n}\frac{1}{2\theta_2} + \sum_{i=1}^{n}\frac{(x_i-\theta_1)^2}{2\theta_2^2} = 0 \; .$$

This leads to the following MLE solution:

$$\hat{\theta}_{\text{MLE},1} = \frac{1}{n}\sum_{i=1}^{n} x_i \approx \mu \; , \qquad \text{(sample mean)}$$

$$\hat{\theta}_{\text{MLE},2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\theta}_{\text{MLE},1})^2 \approx \sigma^2 \; . \qquad \text{(sample variance)}$$

Additional Slide: An Interpretation of MLE (1)

Suppose that each sample of dataset $\mathcal{D} = \{x_i\}_{i=1}^n$ is drawn independently from an underlying distribution $p(x \mid \theta)$, i.e.,

- Empirical distribution $\tilde{p}(x) = \frac{1}{n}\sum_{i=1}^n \delta(x - x_i)$ and model $p(x \mid \theta)$, where $\delta(\cdot)$ is Dirac-delta function
- Direc-delta function has the following characteristics[6]:
  - $\delta(x) = \begin{cases} \infty, & x = 0 \\ 0, & x \neq 0 \end{cases}$
  - $\int_{-\infty}^{\infty} \delta(x) dx = 1$

[6] The Dirac delta is not a function in the traditional sense.

Additional Slide: An Interpretation of MLE (2)

- Model fitting can be done by minimizing a distance between the empirical distribution and model.
- A famous distance Kullback-Leibler (KL) divergence:

$$\mathrm{KL}\big(p \| q\big) \;:=\; \int p(x) \log \frac{p(x)}{q(x)} dx$$

- When KL divergence is selected, we have the correspondence between MLE and KL matching

$$\arg\min_{\theta} \mathrm{KL}\big(\tilde{p} \| p_\theta\big) = \hat{\theta}_{MLE} \; .$$

Proof of “MLE = KL Matching”

- Empirical distribution: $\tilde{p}(x) = \frac{1}{n}\sum_{i=1}^{n} \delta(x - x_i)$
- Model: $p(x \mid \theta) = p_\theta(x)$

$$
\begin{aligned}
\arg\min_\theta \mathrm{KL}\big(\tilde{p} \| p_\theta\big) &= \arg\min_\theta \int \tilde{p}(x) \log \frac{\tilde{p}(x)}{p_\theta(x)} dx \\
&= \arg\min_\theta \left[-H(\tilde{p}) - \int \tilde{p}(x) \log p_\theta(x) dx \right] \\
&= \arg\max_\theta \frac{1}{n} \int \sum_{i=1}^{n} \delta(x - x_i) \log p_\theta(x) dx \\
&= \arg\max_\theta \frac{1}{n} \sum_{i=1}^{n} \log p_\theta(x_i) \;\; =: \;\; \hat{\theta}_{\mathrm{MLE}}
\end{aligned}
$$

Table of Contents

Motivation

We have a coin, and a probability to get a head of the coin is μ.

- We flipped the coin 10 times and observed 7 heads and 3 tails.
- Q: what would be the most plausible value μ given these observations?
    - A: $\hat{\mu}_{\text{MLE}} = 0.7$
- However, we know that a coin is fair in general (i.e. $\mu = 0.5$). So, the result from MLE may be just because of the small number of experiments.
    - How can we encode such belief (a coin is fair) into our statistical framework?

Maximum A Posteriori (MAP)

- As MLE does, MAP has a probability model $p(\mathcal{D} \mid \theta)$ generating data $\mathcal{D}$ from parameter θ; but assumes a priori distribution $p(\theta \mid \alpha)$ of parameter additionally.
  - The hyper-parameter α defines the prior.
  - The Latin phrases: "a priori" = "from the earlier" and "a posteriori" = "from the later"
- MAP finds the parameters $\hat{\theta}_{\mathsf{MAP}}$ maximizing a posteriori distribution $p(\theta \mid \mathcal{D})$, i.e.,

$$\hat{\theta}_{\mathsf{MAP}} := \arg\max_{\theta} p(\theta \mid \mathcal{D})$$

MAP vs. MLE

$$
\begin{aligned}
\hat{\theta}_{\text{MAP}} &:= \arg\max_{\theta} p(\theta \mid \mathcal{D}) \\
&= \arg\max_{\theta} \frac{p(\mathcal{D} \mid \theta) p(\theta)}{p(\mathcal{D})} \\
&= \arg\max_{\theta} p(\mathcal{D} \mid \theta) p(\theta) \\
&= \arg\max_{\theta} \left[\log p(\mathcal{D} \mid \theta) + \log p(\theta)\right] \ .
\end{aligned}
$$

- The prior $p(\theta)$ plays a critical role in protecting against overfitting.
- If our belief says the function should be smooth, then the prior plays like an regularizer, which penalizes too complex models, and values simple ones.

An Example of MAP: Beta-Binomial (0)

Recap the beta distribution

- Beta distribution is a distribution over $[0, 1]$.
- p.d.f, mean, and variance of $\text{Beta}(\mu|\alpha, \beta)$ are $(\alpha, \beta > 0)$

$$p(\mu|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}\mu^{\alpha-1}(1 - \mu)^{\beta-1}$$
$$\mathbb{E}[\mu] = \frac{\alpha}{\alpha + \beta}$$

where $\Gamma(\cdot)$ is a gamma function

$$\Gamma(t) := \int_0^\infty x^{t-1} \exp(-x) dx, \qquad t > 0$$
$$\Gamma(t + 1) = t\Gamma(t)$$

An Example of MAP: Beta-Binomial (1)

Assume we have observed x heads out of n trials of a coin flip from $\text{Bin}(x|\mu, n)$ with unknown μ. Use a prior $\text{Beta}(\mu|\alpha, \beta)$. Then, MAP solution maximizes the following loss function:

$$
\begin{aligned}
\mathcal{L}_{MAP}(\mu) &= \log p(x|\mu) + \log p(\mu) \\
&= \log \binom{n}{x} \mu^{x}(1-\mu)^{n-x} + \log \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \mu^{\alpha-1}(1-\mu)^{\beta-1} \\
&\propto x \log \mu + (n-x)\log(1-\mu) \\
&\quad + (\alpha-1)\log \mu + (\beta-1)\log(1-\mu)
\end{aligned}
$$

Then, it follows from solving $\frac{\partial \mathcal{L}_{\text{MAP}}}{\partial \mu} = 0$ that

$$
\hat{\mu}_{\text{MAP}} = \frac{\alpha + x - 1}{\alpha + \beta + n - 2} \, .
$$

An Example of MAP: Beta-Binomial (2)

$$\hat{\mu}_{\mathrm{MAP}} = \frac{\alpha + x - 1}{\alpha + \beta + n - 2}$$

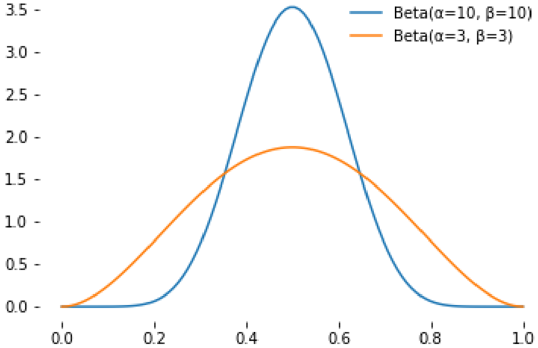
- Let $\alpha = \beta = 3$, $x = 7$ and $n = 10$, then

$$\hat{\mu}_{\mathrm{MAP}} = \frac{9}{14} = 0.64 \cdots < 0.7 = \hat{\mu}_{\mathrm{MLE}}$$

- α and β is our prior belief about the fairness of a coin.
- As we increases α and β, $\hat{\mu}_{\mathrm{MAP}}$ approaches to one half.
- In case of $n \gg \alpha + \beta$, i.e., prior is weaker than data, we have

$$\hat{\mu}_{\mathrm{MAP}} \simeq \hat{\mu}_{\mathrm{MLE}}$$

An Example of MAP: Beta-Binomail (3)



Beta distributions with two parameters α and β.

An Example of MAP: Gaussian (1)

Assume $\mathcal{D}$ is n i.i.d. copies of univariate Gaussian random variable $\mathcal{N}(\mu, 1)$ with unknown[7] μ. Use a prior $p(\mu \mid \alpha) \sim \mathcal{N}(0, \alpha^2)$. Then, MAP solution maximizes the following loss function:

$$\begin{aligned}
\mathcal{L}_{MAP}(\theta) &= \log p(\mathcal{D} \mid \theta) + \log p(\theta) \\
&\propto \left[-\frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^2 - \frac{1}{2\alpha^2} \mu^2 \right]
\end{aligned}$$

Then, it follows from solving $\frac{\partial \mathcal{L}_{\text{MAP}}}{\partial \mu} = 0$ that

$$\hat{\mu}_{\text{MAP}} = \frac{1}{\left(n + \frac{1}{\alpha^2}\right)} \sum_{i=1}^{n} x_i \; .$$

[7] Here we assume that we know the variance.

An Example of MAP: Gaussian (2)

- In case of $n \gg \frac{1}{\alpha^2}$, i.e., prior is weaker than data, we have

$$\hat{\mu}_{\mathrm{MAP}} \quad \simeq \quad \hat{\mu}_{\mathrm{MLE}} = \frac{1}{n} \sum_{i=1}^{n} x_i \ .$$

- In case of $n \ll \frac{1}{\alpha^2}$, i.e., prior is stronger than data, we have

$$\hat{\mu}_{\mathrm{MAP}} \quad \simeq \quad 0 \ .$$

If only few data points are available, the prior will bias the estimate towards the priori expected value.

Table of Contents

Motivation

We have a coin, and a probability to get a head of the coin is μ.

- We flipped the coin 10 times and observed 7 heads and 3 tails.
- Q: what would be the most plausible value μ given these observations?
  - A: $\hat{\mu}_{\text{MLE}} = 0.7$ with MLE and $\hat{\mu}_{\text{MAP}} < 0.7$ with MAP.
- However, would it be okay to represents the results as a single number?
  - How much are we sure about the results? (uncertainty)

MLE/MAP as Point-wise Estimator

MLE/MAP extracts a value of parameter $\hat{\theta} = \hat{\theta}_{\text{MLE}}$ or $\hat{\theta}_{\text{MAP}}$ representing dataset $\mathcal{D}$. From which, our prediction can be done via

- Unsupervised $p(x_{\text{new}} \mid \mathcal{D}; \alpha)$ would be $p(x_{\text{new}} \mid \hat{\theta})$.
- Supervised $p(y_{\text{new}} \mid x_{\text{new}}, \mathcal{D}; \alpha)$ would be $p(y_{\text{new}} \mid x_{\text{new}}, \hat{\theta})$.

Again the prediction is made on a single estimated value.

Due to this property, we call MLE/MAP as a point-wise estimator.

MLE/MAP vs Bayesian Inference

Bayesian inference tries to estimate them directly via a weighted average over all values of θ instead of choosing a specific value of parameter:

- Unsupervised Bayesian

$$
\begin{aligned}
p(x_{\text{new}} \mid \mathcal{D}; \alpha) &= \int p(x_{\text{new}} \mid \theta, \mathcal{D}; \alpha) p(\theta \mid \mathcal{D}; \alpha) d\theta \\
&= \int p(x_{\text{new}} \mid \theta) \underbrace{p(\theta \mid \mathcal{D}; \alpha)}_{\text{MAP}} d\theta \, .
\end{aligned}
$$

- Supervised Bayesian

$$
\begin{aligned}
p(y_{\text{new}} \mid x_{\text{new}}, \mathcal{D}; \alpha) &= \int p(y_{\text{new}} \mid x_{\text{new}}, \theta, \mathcal{D}; \alpha) p(\theta \mid \mathcal{D}; \alpha) d\theta \\
&= \int p(y_{\text{new}} \mid x_{\text{new}}, \theta) p(\theta \mid \mathcal{D}; \alpha) d\theta \, .
\end{aligned}
$$

Therefore, we need a posterior distribution! (instead of a point that maximize the posterior)

Bayesian Inference: Posterior Calculation

The posterior distribution of θ is updated using Bayes rule, where the likelihood is given by $p(\mathcal{D} \mid \theta) = \prod_{i=1}^{n} p(x_i \mid \theta)$:

$$\begin{aligned} p(\theta \mid \mathcal{D}) &= \frac{p(\mathcal{D} \mid \theta) p(\theta)}{p(\mathcal{D})} \\ &= \frac{p(\theta) \prod_{i=1}^{n} p(x_i \mid \theta)}{\underbrace{\int p(\theta') \prod_{i=1}^{n} p(x_i \mid \theta') d\theta'}_{\text{We don't ignore anymore}}} \end{aligned}$$

Conjugate prior: a good choice of prior for the ease of analysis

- A prior $p(\theta)$ which gives rise to a posterior $p(\theta \mid \mathcal{D})$ having the same function form, given $p(\mathcal{D} \mid \theta)$.

Some Conjugate Priors[8]

Prior $p(\theta \mid \alpha)$	Likelihood $p(\mathcal{D} \mid \theta)$	Posterior $p(\theta \mid \mathcal{D}, \alpha)$
Beta	Benoulli	Beta
Beta	Binomial	Beta
Normal	Normal	Normal
Gamma	Gamma	Gamma
Gamma	Poisson	Gamma
Normal-Gamma	Normal	Normal-Gamma

[8] https://en.wikipedia.org/wiki/Conjugate_prior#Table_of_conjugate_distributions

Beta-Bernoulli Conjugacy (1)

Think about coin toss with observation x.

- The likelihood of observing x can be modeled with Bernoulli parameterized by μ, i.e, $p(x|\mu) = \mathrm{Ber}(x|\mu)$.
- We don't know μ, but we can place a Beta distribution parmeterized by α, β,. i.e $p(\mu|\alpha, \beta) = \mathrm{Beta}(\mu|\alpha, \beta)$.
- Bayes rule tells us the posterior of μ given x as

$$p(\mu|x, \alpha, \beta) = \frac{p(x|\mu)p(\mu|\alpha, \beta)}{p(x|\alpha, \beta)}$$

- Compute the posterior!

Beta-Bernoulli Conjugacy (2)

The marginal $p(x \mid \alpha, \beta)$ can be obtained by

$$
\begin{aligned}
p(x \mid \alpha, \beta) &= \int p(x|\mu)p(\mu|\alpha,\beta)d\mu \\
&= \int \mu^{x}(1-\mu)^{1-x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\mu^{\alpha-1}(1-\mu)^{\beta-1}d\mu \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\underbrace{\int \mu^{x+\alpha-1}(1-\mu)^{\beta-x}d\mu}_{\text{Beta function}} \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+x)\Gamma(\beta-x+1)}{\Gamma(\alpha+\beta+1)}
\end{aligned}
$$

The posterior distribution can then be derived as

$$
\frac{p(x|\mu)p(\mu|\alpha,\beta)}{p(x|\alpha,\beta)} \sim \mathrm{Beta}(x+\alpha, \beta-x+1)
$$

Beta-Bernoulli Conjugacy (3)

Since $p(x|\alpha, \beta)$ is just a part of normalizing constant making $\int p(\mu|x)d\mu = 1$, you can directly obtain posterior from

$$
\begin{aligned}
\frac{p(x|\mu)p(\mu|\alpha,\beta)}{p(x|\alpha,\beta)} &\propto \mu^x(1-\mu)^{1-x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\mu^{\alpha-1}(1-\mu)^{\beta-1} \\
&\propto \mu^{x+\alpha-1}(1-\mu)^{\beta-x} \\
&\sim \text{Beta}(x+\alpha, \beta-x+1)
\end{aligned}
$$

In other words, from $\int A\mu^{x+\alpha-1}(1-\mu)^{\beta-x}d\mu = 1$ where A is a normalizing constant, we can directly obtain Beta distribution.

This result can be generalized to the Beta-Binomial case.

Some Conjugate Priors[9]

Prior $p(\theta \mid \alpha)$	Likelihood $p(\mathcal{D} \mid \theta)$	Posterior $p(\theta \mid \mathcal{D}, \alpha)$
Beta	Benoulli	Beta
Beta	Binomial	Beta
Normal	Normal	Normal
Gamma	Gamma	Gamma
Gamma	Poisson	Gamma
Normal-Gamma	Normal	Normal-Gamma

[9] https://en.wikipedia.org/wiki/Conjugate_prior#Table_of_conjugate_distributions

Bayesian Inference: Normal-Normal (1)

For a given set $\mathcal{D} = \{x_i\}_{i=1}^{n}$ of n real numbers, assume that:

- (as model) each x_i is drawn independently from $\mathcal{N}(\mu, \sigma^2)$
- (as prior) σ^2 is known in advance, and μ is drawn from $\mathcal{N}(\mu_0, \sigma_0^2)$, of which density function is denoted by $p_0(\mu; \mu_0, \sigma_0^2)$.
  The posterior is calculated as follows:

$$p(\mu \mid \mathcal{D}) = \frac{p_0(\mu)}{p(\mathcal{D})} \prod_{i=1}^{n} p(x_i \mid \mu) \ ,$$

  where

$$p(x_i \mid \mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (x_i - \mu)^2 \right) \ .$$

Bayesian Inference: Normal-Normal (2)

After a basic calculus, we have

$$p(\mu \mid \mathcal{D}) = \frac{1}{\sqrt{2\pi\tilde{\sigma}^2}} \exp\left(-\frac{1}{2\tilde{\sigma}^2}(\mu - \tilde{\mu})^2\right) ,$$

where

$$\tilde{\mu} = \frac{\frac{\mu_0}{\sigma_0^2} + \sum_{i=1}^{n} \frac{1}{\sigma^2} x_i}{\frac{1}{\sigma_0^2} + \sum_{i=1}^{n} \frac{1}{\sigma^2}} \quad \text{and} \quad \frac{1}{\tilde{\sigma}^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} .$$

- When $n = 0$, $\tilde{\mu}$ reduces to the prior mean μ.
- As $n \to \infty$, the posterior mean is given by the ML solution.

Additional Reading

- Section 3 of the text book (Probabilistic Machine Learning: An Introduction)
- Supplementary material on PLMS (Bayesian_Normal.pdf)