Week-4 | Summary

Karthik Thiagarajan

1. Common

1.1. Notation

Scalars:

$$x_1, x_2, y_1, y_2, z_2, z_2, a, b, \alpha, \beta$$

Column vector:

$$\mathbf{x} \in \mathbb{R}^d$$

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$$

Row vector:

$$\mathbf{x} \in \mathbb{R}^d$$

$$\mathbf{x}^T = \begin{bmatrix} x_1 & \cdots & x_d \end{bmatrix}$$

Matrix:

$$\mathbf{X} \in \mathbb{R}^{d \times n}$$

1.2. Dataset

$$D = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$$

1.3. Data-matrix

$$\mathbf{X} \in \mathbb{R}^{d \times n}$$

- $d \rightarrow$ number of features
- $n \rightarrow$ number of data-points

$$X = \begin{bmatrix} | & & | \\ \mathbf{x}_1 & \cdots & \mathbf{x}_n \\ | & & | \end{bmatrix}$$

1.4. Data-point

$$\mathbf{x}_i \in \mathbb{R}^d$$

2. Estimation: MLE

Likelihood

The likelihood of a dataset D under a distribution parameterized by θ is given below:

$$L(\theta; D) = \prod_{i=1}^{n} P(\mathbf{x}_i; \theta)$$

- The likelihood is the "likelihood" of seeing the data if it is the result of drawing samples from the underlying distribution.
- It takes this particular form if the points are assumed to be sampled independently and identically from the distribution.
- The likelihood is a function of the parameter θ. It should not be confused with a probability distribution.
- P could be a PDF or a PMF depending on the whether $\mathbf{x}_i$ is discrete or continuous.

Log-likelihood

$$l(\theta; D) = \sum_{i=1}^{n} \log P(\mathbf{x}_i; \theta)$$

Since product of probabilities would result in a very small number, we move to log-space to avoid underflow.

Maximizing the likelihood

Estimate the parameter value that maximizes the likelihood:

$$\max_{\theta} \; L(\theta; D)$$

Maximizing the log-likelihood

Since $\log$ is a strictly increasing function, we can maximize the log-likelihood instead:

$$\max_{\theta} \; l(\theta; D)$$

2.1. Example: Bernoulli

Support

$$\{0, 1\}$$

$X = 1$ is equivalent to heads and $X = 0$ is equivalent to tails.

PMF

$$P(X = x) = p^x (1 - p)^{1-x}$$

A compact representation of $P(X = 1) = p$ and $P(X = 0) = 1 - p$.

Likelihood

$$L(p; D) = p^{\sum_{i=1}^{n} x_i} \; (1 - p)^{\sum_{i=1}^{n} (1 - x_i)}$$

Simplifies to

$$L(p; D) = p^{n_h}(1-p)^{n_t}$$

where n_h is number of heads and n_t is number of tails. Note $n_h + n_t = n$.

Log-likelihood

$$l(p; D) = n_h \log p + n_t \log(1-p)$$

MLE for p

$$\widehat{p} = \frac{n_h}{n}$$

2.2. Gaussian

Support

$$\mathbb{R}$$

PDF

$$f\big(x; \mu, \sigma^2\big) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

Likelihood

$$L\big(\mu, \sigma^2; D\big) = \prod_{i=1}^{n} f\big(x_i; \mu, \sigma^2\big)$$

Log-likelihood

$$L\big(\mu, \sigma^2; D\big) = \sum_{i=1}^{n} \log f\big(x_i; \mu, \sigma^2\big)$$

MLE for μ

$$\widehat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}$$

3. Estimation: Bayesian methods

In a Bayesian setting, probabilities are viewed as beliefs.

Bayes Theorem

The Bayes theorem is a tool that allows you to update your belief about a situation using data.

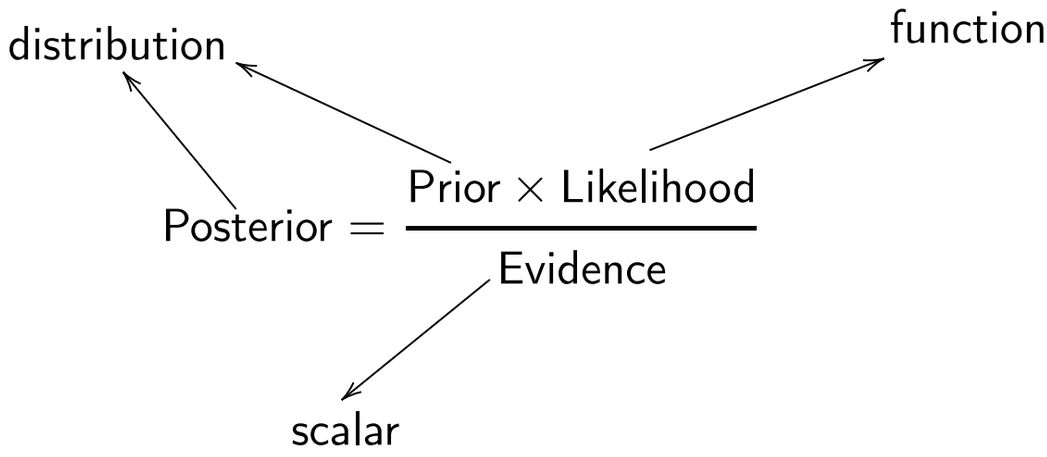
$$\text{Posterior} = \frac{\text{Prior} \times \text{Likelihood}}{\text{Evidence}}$$

- The prior encodes your prior belief about the situation before observing the data (evidence).
- The likelihood tells you how well the data conforms to your prior belief.
- The likelihood is multiplied with the prior and normalized with the evidence to give the posterior, the updated belief.
- The evidence is a normalizing factor here.

In the context of parameter estimation, Bayes theorem takes this form:

$$P(\theta \mid D) = \frac{P(\theta) \cdot P(D \mid \theta)}{P(D)}$$

A note on the type of objects in the Bayes theorem:



We will look at an example of Bayesian estimation for a binary dataset in $\{0, 1\}^n$ modeled using a Bernoulli distribution with a Beta prior.

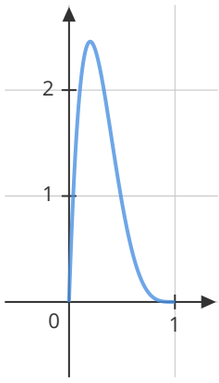
3.1. Bernoulli with Beta prior

Prior

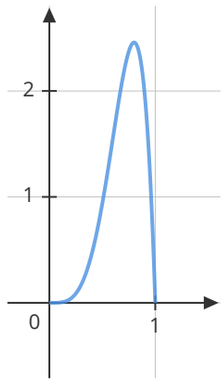
$$\text{Beta}(\alpha, \beta) = \frac{1}{B(\alpha, \beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

- $\alpha, \beta > 0$ are parameters of the distribution
- Support is $[0, 1]$.
- $B(\alpha, \beta)$ is a normalizing constant that ensures that f is a PDF

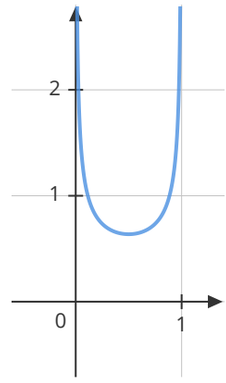
A quick look at some of the possible shapes of the Beta distribution. Each one can model a different



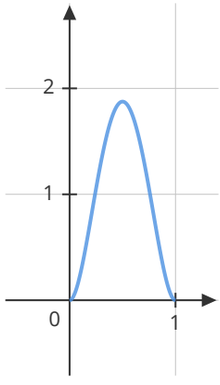
Beta$(2, 5)$



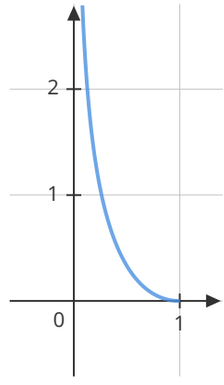
Beta$(5, 2)$



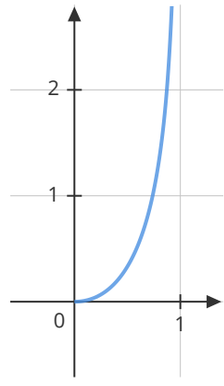
Beta$(0.5, 0.5)$



Beta$(3, 3)$



Beta$(0.5, 3)$



Beta$(3, 0.5)$

Likelihood

The Bernoulli likelihood:

$$p^{n_h}(1-p)^{n_t}$$

Posterior

$$\text{Posterior} \propto \text{Prior} \times \text{Likelihood}$$

$$\text{Posterior} \propto p^{\alpha-1}(1-p)^{\beta-1}p^{n_h}(1-p)^{n_t}$$

$$\text{Posterior} = \text{Beta}(\alpha + n_h, \beta + n_t)$$

The Beta distribution is a conjugate prior for the Bernoulli likelihood. A

conjugate prior has a similar form as the likelihood simplifying the computation of the posterior.

3.2. Point estimate

Often we would want a point estimate (a single number) for the parameter. But Bayesian methods return a distribution over the parameter. We look at two ways to extract a point estimate:

- expectation of the posterior
- mode of the posterior

For this example, the expected value of the posterior is:

$$\frac{\alpha + n_h}{\alpha + \beta + n}$$

The mode of the posterior for $\alpha + n_h > 1, \beta + n_t > 1$ is:

$$\frac{\alpha + n_h - 1}{\alpha + \beta + n - 2}$$

The mode of the posterior is often called the Maximum A Posteriori estimate or MAP estimate, since the mode is nothing but the (arg)maximum of the posterior.

4. Gaussian Mixture Models

For more complex distributions, we have what is called a Gaussian Mixture Model. A GMM is a probability distribution. It is a mixture of K Gaussians, each of which is called a component.

$$f(x; \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(x; \mu_k, \sigma_k^2\right)$$

where $\sum_{k=1}^{n} \pi_k = 1$ so that f is a valid PDF.

- f is the PDF of the GMM.
- $\mathcal{N}$ is the PDF of a Gaussian distribution with mean μ_k and variance σ_k^2
- π_k is the "prior" contribution of the k^{th} component and are called the mixture probabilities.

A GMM is a latent variable model. That is, we can view the data-generation process by introducing a latent (hidden) variable z_i for each data-point. Generating x_i can be explained as follows:

- First choose a component k by setting $z_i = k$ with prior probability π_k
- Sample a point from this Gaussian; the conditional density associated with this is $\mathcal{N}\left(x; \mu_k, \sigma_k^2\right)$

The joint density of seeing the point x_i from component k becomes:

$$f(X = x_i, Z = k) = \pi_k \mathcal{N}\left(x_i; \mu_k, \sigma_k^2\right)$$

Marginalizing over the random variable Z would give us the density:

$$f(X = x_i) = \sum_{k=1}^{K} f(X = x, Z = k)$$

Leading us to:

$$f(x_i) = \sum_{k=1}^{K} \pi_k \cdot \mathcal{N}\left(x_i; \mu_k, \sigma_k^2\right)$$

This is the density of the GMM as seen before but explained using latent variables. Note that the latent variable is not explicitly observed. We posit that such a variable exists. Only the dataset D is observed.

For a GMM with K components, we need to estimate $3K$ parameters:

- K mixture probabilities
- K means
- K variances

5. EM algorithm

We can use MLE to estimate the parameters. But we don't have a closed form solution. Thankfully, we have an iterative approach to parameter estimation called the EM algorithm. This algorithm makes use of some intermediate variables that help in parameter estimation:

$$\lambda_k^i$$

Points to note:

- i: corresponds to index of the data-point
- k: corresponds to index of the component
- λ_k^i can be interpreted as a conditional probability
  - $0 \leqslant \lambda_k^i \leqslant 1$
  - $\sum_{k=1}^{n} \lambda_k^i = 1$ for all i
- There are KN such variables

The parameters are collectively referred to as $\boldsymbol{\theta} = [\pi, \mu, \sigma]$. We keep bettering our estimate of $\boldsymbol{\theta}$ in each step. There are two steps in the algorithm:

- E-step: update the values for $\boldsymbol{\lambda}$ using the current values of $\boldsymbol{\theta}$
- M-step: update the values of $\boldsymbol{\theta}$ using the newly found values of $\boldsymbol{\lambda}$

Convergence criterion

When successive iterates become smaller than some ϵ

$$||\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}|| < \epsilon$$

Initialization

Use K-means algorithm to initialize $\theta_k = [\pi_k, \mu_k, \sigma_k]$.

Until convergence

E-step

$$\lambda_k^i = P(z_i = k \mid X = x_i)$$

λ_k^i is the contribution of the k^{th} component to the point x_i given that we have observed the point x_i. It represents the posterior probability of Z given X, that is, $P(Z \mid X)$.

Using the Bayes' theorem:

$$\lambda_k^i = \frac{P(z_i = k) \cdot P(X = x_i \mid z_i = k)}{P(X = x_i)}$$

$$\pi_k \cdot \mathcal{N}\left(x_i; \mu_k, \sigma_k^2\right)$$

$$= \frac{}{\sum_{j=1}^{k} \pi_j \cdot \mathcal{N}\left(x_i; \mu_j, \sigma_j^2\right)}$$

Here we use the current values of π_k, μ_k, σ_k^2 to estimate λ_k^i.

M-Step

Use the values of λ_k^i obtained in the E-step to update the values of π_k, μ_k, σ_k^2.

$$\mu_k = \frac{\sum_{i=1}^{n} \lambda_k^i \cdot x_i}{\sum_{i=1}^{n} \lambda_k^i}, \qquad \sigma_k^2 = \frac{\sum_{i=1}^{n} \lambda_k^i \cdot (x_i - \mu_k)^2}{\sum_{i=1}^{n} \lambda_k^i}$$

$$\pi_k = \frac{\sum_{i=1}^{n} \lambda_k^i}{n}$$

Soft clustering

EM algorithm can be seen as method that does soft-clustering. λ_k^i can be seen as the affinity of x_i to component k. In K-means this affinity is binary -- a point belongs to a cluster or not. In the case of EM, this affinity is a number between $[0, 1]$. The E-step is analogous to the cluster assignment step in K-means. The M-step is analogous to the updates for the cluster centers in K-means.