2022

Lecture 7: Linear Regression

Week 7

Lecturer: Tianyu Wang

1 Basics of Bayesian Analysis

The Bayesian rule is a simple but surprisingly useful fact. For two events A and B such that $\mathbb{P}(B) \neq 0$, it holds that

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

The proof follows from simple manipulation of the definition of conditional probability:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}.$$

Now let's assign A and B more specific meanings. Let A be the belief about the model, and let B the the data. We have

$$\mathbb{P}(''model''|''data'') = \frac{\mathbb{P}(''data''|''model'')\mathbb{P}(''model'')}{\mathbb{P}(''data'')}.$$

In the above, $\mathbb{P}(''data''|''model'')$ is called likelihood, $\mathbb{P}(''model'')$ is called prior, and $\mathbb{P}(''model''|''data'')$ is called posterior. Usually, $\mathbb{P}(''data'')$ is treated as a normalizing constant so that $\int_{''model''} \mathbb{P}(''model''|''data'') = 1$, and the Bayes rule for statistical inference is written as

$$\mathbb{P}(''model''|''data'') \propto \mathbb{P}(''data''|''model'')\mathbb{P}(''model'').$$

1.1 Examples

Suppose we want to estimate the probability of Bernoulli distribution X. We suppose that $\mathbb{P}(X = 1) = \theta$. Let's say we have a prior belief about θ, which follows a beta distribution. The beta distribution is parametrized by two parameters a and b. The density for beta distribution $Beta(a, b)$ is $f(x) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}$ for $x \in [0, 1]$, where $B(a, b) = \int_0^1 x^{a-1}(1 - x)^{b-1}\, dx$ is a normalization constant. $B(a, b)$ is called the beta function. The expectation of a random variable from beta distribution $Beta(a, b)$ is $\frac{a}{a+b}$. The density function of some beta distributions are in Figure 1.

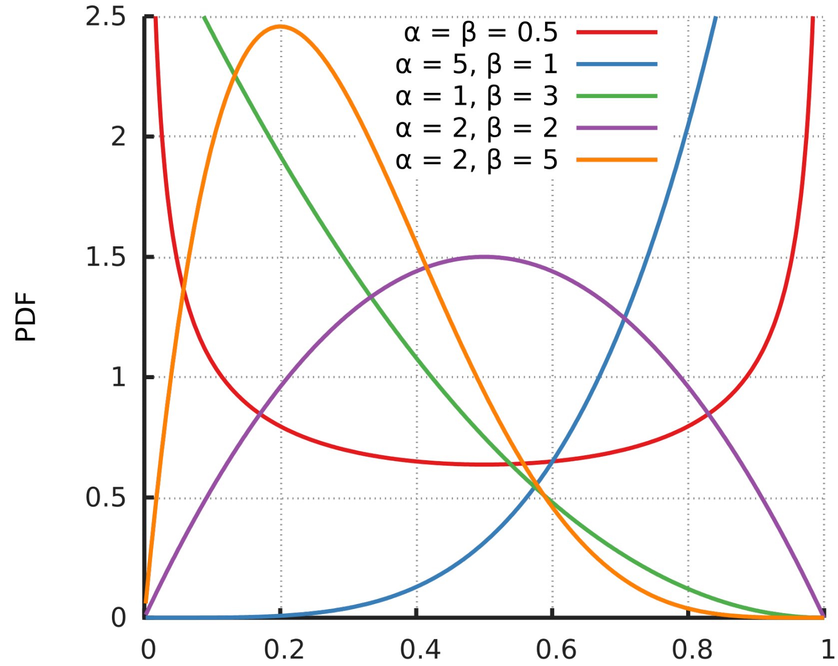

Figure 1: Density function of some beta distributions. Source:wikipedia.

Let's say our prior distribution about θ is that $\theta \sim Beta(a, b)$ for some constants a, b. After observing $X_1, X_2, X_3, \cdots, X_n$ $i.i.d.$ samples from the Bernoulli distribution, our posterior belief about θ becomes

$$\begin{aligned}\mathbb{P}(\theta|X_1, X_2, \cdots, X_n) &\propto \mathbb{P}(X_1, X_2, \cdots, X_n|\theta)\mathbb{P}(\theta) \\ &= \prod_{i=1}^{n} \theta^{X_i}(1-\theta)^{1-X_i}\theta^{a-1}(1-\theta)^{b-1} \\ &= \theta^{a+\sum_{i=1}^{n} X_i - 1}(1-\theta)^{b+\sum_{i=1}^{n}(1-X_i)-1}.\end{aligned}$$

This is the beta distribution with parameters $a + \sum_{i=1}^{n} X_i$ and $b + \sum_{i=1}^{n}(1 - X_i)$. Note that the expectation of this posterior distribution is $\frac{a+\sum_{i=1}^{n} X_i}{a+b+n}$, which is $a +$ "number of heads" divided by $a + b +$ "total number of coin tosses".

Conjugate Prior

The definition of conjugate prior is given below:

> If the posterior distribution is in the same probability distribution family as the prior probability distribution, the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior for the likelihood function.

In the above example, we see that beta distribution is a conjugate prior for Bernoulli likelihood.

Maximum-A-Posterior Estimator

Note that the above procedure gives a distribution over the model parameters θ. One way to exact an estimator for the model parameter θ is to maximize over the posterior. This gives the Maximum-A-Posterior (MAP) estimator:

$$\widehat{\theta}^{MAP} \in \arg\max_{\theta} \mathbb{P}(\theta | X_1, X_2, \cdots, X_n) = \arg\max_{\theta} \mathbb{P}(X_1, X_2, \cdots, X_n | \theta) \mathbb{P}(\theta) .$$

There will be question(s) on Bayesian inference in the homework.

2 Bayesian Linear Regression

In linear regression, we want to find $\theta \in \mathbb{R}^d$, so that $f(x) = \theta^\top x$ fits a dataset $\{(x_i, y_i)\}_{i=1}^n$.

Let θ follow a standard Gaussian prior: $\mathbb{P}(\theta) = N(0, \lambda I)$, where I is the $d \times d$ identity matrix and λ is a positive constant. Consider the likelihood model $Y|X \sim N(\theta^\top X, 1)$. Then the posterior is

$$\begin{aligned} \mathbb{P}\left(\theta | \{(x_i, y_i)\}_{i=1}^n\right) &\propto \mathbb{P}\left(\{y_i\}_{i=1}^n | \{x_i\}_{i=1}^n, \theta\right) \mathbb{P}(\theta) \\ &\propto \prod_{i=1}^n \exp\left(-\frac{\left(\theta^\top x_i - y_i\right)^2}{2}\right) \exp\left(-\frac{\lambda \|\theta\|_2^2}{2}\right) . \end{aligned}$$

Let's look at what the MAP estimator gives:

$$\theta^{MAP} \in \arg\max_{\theta} \left\{ \prod_{i=1}^n \exp\left(-\frac{\left(\theta^\top x_i - y_i\right)^2}{2}\right) \exp\left(-\frac{\lambda \|\theta\|_2^2}{2}\right) \right\} .$$

Compare the MAP estimator from this Bayesian linear regression model to solution of the ridge regression problem.

3 Kernelized Linear Regression & Gaussian Processes

Continue next time...

Acknowledgement

Reference: Machine Learning: A Probabilistic Perspective by Kevin Murphy. A thank you to wikipedia contributors.