CS 337, Fall 2023
Probabilistic Model of Linear Regression, MLE, MAP, Conjugate Priors

Scribes: Manasi Ingle, Sreelaxmi Gasada, Kanishk Garg,
Modulla Hrushikesh Reddy, Ankan Sarkar, Shantanu Welling*
Edited by: Bhavani Shankar

Monday 14th August, 2023

1 Probabilistic Model Of Linear Regression

The probabilistic model of linear regression provides a way to understand and interpret linear regression models from a probabilistic view.

Consider training data $\mathcal{D}$:

$$\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$$

Assumption: The relation between the input features $\mathbf{x}$ and the target y can be represented by a linear equation with some added noise ϵ:

$$y_i = \mathbf{w}^T \mathbf{x}_i + \epsilon_i$$

Noise: The noise component ϵ signifies the fluctuation within the data that contributes towards uncertainty in the target values and cannot be explained by the linear relationship. Assume that it follows a Gaussian distribution with mean 0 and variance σ^2

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

In other words, ϵ_i's are independent and identically distributed (i.i.d.) Gaussian random variables with zero-mean and the same variance σ^2. Now, we can represent target y_i as a Gaussian distribution with mean $\mathbf{w}^T \mathbf{x}_i$ and variance σ^2

$$y_i \sim \mathcal{N}(\mathbf{w}^T \mathbf{x}_i, \sigma^2)$$

Given training data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ where feature vector $\mathbf{x}_i \in \mathbb{R}^d$, target $y_i \in \mathbb{R}$,

$$P(y_i \mid \mathbf{x}_i, \mathbf{w}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i - \mathbf{w}^T \mathbf{x}_i)^2}{2\sigma^2}}$$

*All scribes get equal credit for the final notes.

$$\text{Likelihood}: \quad P(y_1, y_2, \ldots, y_n \mid \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) = \pi_i P(y_i \mid \mathbf{x}_i, \mathbf{w})$$

$$\text{Log Likelihood}: \quad \log P(y_1, y_2, \ldots, y_n \mid \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) = \sum_i \log P(y_i \mid \mathbf{x}_i, \mathbf{w})$$

The likelihood function models how well the observed data fits the assumed model. In case of linear regression, it measures the probability of observing the actual target values given the parameter vector w and feature vectors.

2 Maximum Likelihood Estimation

The process of finding the best-fitting model often involves estimating the model's parameters in a way that maximizes the likelihood of observing the actual data. This estimation process is known as Maximum Likelihood Estimation (MLE). It is a *Frequentist* view in machine learning.

For the model, y_i(target value) represented as a Gaussian distribution $\mathcal{N}(\mathbf{w}^T\mathbf{x}_i, \sigma^2)$, the parameter of interest is the weight vector $\mathbf{w}$. We need to find $\mathbf{w}$ that maximizes the likelihood function. Maximizing likelihood function is equivalent to maximizing log likelihood function. (The log function is strictly increasing, and thus maximizing the log likelihood gives the same solution as maximizing the likelihood function.)

$$\mathbf{w}_{\text{MLE}} = \arg\max_{\mathbf{w}} \sum_i \log P(y_i \mid \mathbf{x}_i, \mathbf{w})$$

2.1 Motivating Example

You want to estimate the probability of a biased coin landing on heads. The probability that the outcome is head for a coin toss is θ. Let us say your data contains N coin tosses with N_H heads and N_T tails. What is the maximum likelihood estimate of θ ?

Likelihood function: probability of data, given parameter θ

$$P(D \mid \theta) = \theta_H^N (1-\theta)_T^N$$

$$\begin{aligned}
\theta_{\text{MLE}} &= \arg\max_{\theta} \log P(\mathcal{D}|\theta) && (1)\\
&= \arg\max_{\theta} \log \left(\theta^{N_H}(1-\theta)^{N_T}\right) && (2)\\
&= \arg\max_{\theta} \left(N_H \log\theta + N_T \log(1-\theta)\right) && (3)
\end{aligned}$$

To find θ_{MLE}, we need to find the derivative of $N_H \log\theta + N_T \log(1-\theta)$ w.r.t. θ, set to zero and solve for θ:

$$\frac{N_H}{\theta_{\text{MLE}}} - \frac{N_T}{1-\theta_{\text{MLE}}} = 0$$

$$\theta_{\text{MLE}} = \frac{N_H}{N_H + N_T} = \frac{N_H}{N}$$

2.2 Question

Say you have a coin with P $\in \{0.4, 0.6\}$
Data: 3 coin tosses; 2 heads, 1 tail
What is MLE of P ?

Solution:

$$P(D \mid \theta)_{\theta=0.4} = (0.4)^2(0.6)$$
$$P(D \mid \theta)_{\theta=0.6} = (0.6)^2(0.4)$$
$$P(D \mid \theta)_{\theta=0.6} > P(D \mid \theta)_{\theta=0.4}$$
$$\textbf{Answer = 0.6}$$

2.3 MLE for linear regression

$$w_{\text{MLE}} = \arg\max \sum_i \log P(y_i \mid \mathbf{x}_i, \mathbf{w}) \tag{4}$$
$$= \arg\max \sum_i \log(\mathcal{N}(\mathbf{w}^T\mathbf{x}_i, \sigma^2)) \tag{5}$$
$$= \arg\max \sum_i \log \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i - \mathbf{w}^T\mathbf{x}_i)^2}{2\sigma^2}} \tag{6}$$
$$= \arg\max C - \sum_i \frac{(y_i - \mathbf{w}^T\mathbf{x}_i)^2}{2\sigma^2} \tag{7}$$
$$w_{\text{MLE}} = \arg\min \sum_i (y_i - \mathbf{w}^T\mathbf{x}_i)^2 \tag{8}$$

- MLE under Gaussian noise distribution is equivalent to the least squares solution.
- MLE under Laplacian noise distribution is equivalent to Least Absolute Deviation solution.
- A significant drawback of Maximum Likelihood Estimation (MLE) arises when working with limited data, as it can lead to overfitting due to its heavy reliance on the available dataset.

3 Bayesian Parameter Estimation

In MLE, observations are random variables while parameters are not. On the contrary, in the Bayesian framework, parameters are also treated as random variables alongside observations. Parameters have an underlying **prior distribution** which encode beliefs prior to observing the data. As MLE has a high tendency to overfit, we switch to a different estimation, namely the **Maximum Aposteriori Estimation** with the hope of alleviating overfitting using prior information.

4 Maximum Aposteriori Estimate (MAP)

Maximum A Posteriori (MAP) estimation focuses on incorporation of prior knowledge into estimating the underlying model parameters which are denoted by θ. We assume a prior distribution $P(\theta)$ for parameters θ (before observing the data D).

By Bayes' theorem, we have :

$$P(\theta|\mathcal{D}) = \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})}$$

$\boldsymbol{P(\theta|\mathcal{D})}$ is the **posterior** probability for the model parameters θ given data $\mathcal{D}$.
$\boldsymbol{P(\mathcal{D}|\theta)}$ is the **likelihood** of observing the data given model parameters & $\boldsymbol{P(\theta)}$ is the **prior**.

$$\theta_{\text{MAP}} = \arg\max_{\theta} P(\theta|\mathcal{D}) \tag{9}$$

$$= \arg\max_{\theta} (\log P(\mathcal{D}|\theta) + \log P(\theta)) \tag{10}$$

Note:

The posterior is directly proportional to the product of likelihood and the prior functions as shown by the following equation,

$$P(\theta|\mathcal{D}) \propto P(\mathcal{D}|\theta)P(\theta)$$

$$\implies \log P(\theta|\mathcal{D}) \propto \log P(\mathcal{D}|\theta) + \log P(\theta)$$

4.1 Coin example (MAP estimate)

Suppose, we have a biased coin and $\mathrm{N_H}$ and $\mathrm{N_T}$ are the number of heads and tails obtained.
Likelihood $\mathrm{P}(\mathcal{D} \mid \theta)$, i.e, the probability of data given the parameter θ is given by the following equation,

$$P(\mathcal{D}|\theta) = \theta_H^N (1-\theta)_T^N$$

What is a good prior on θ ?

Beta distribution is a good candidate for prior because the functional form of prior matches with that of the likelihood.

$$Beta(\theta; \alpha, \beta) = c\,\theta^{\alpha-1}(1-\theta)^{\beta-1} \qquad [c = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)},\ \Gamma(.) \text{ is the gamma function}]$$

5 Conjugate Priors

For a likelihood $\mathrm{P}(\mathcal{D} \mid \theta)$ coming from a family of distributions d_1, a prior (from a family d_2) is said to be a conjugate prior if the posterior distribution also comes from the family d_2. (d_2 could also be from the same family as d_1).

5.1 Coin example

Continuing with the coin example mentioned above, the probability of getting $\mathrm{N_H}$ heads and $\mathrm{N_T}$ tails upon tossing a biased coin with the probability of getting head as θ (Bernoulli distribution), is given by,

$$B(N_{\mathrm{H}}, N_{\mathrm{T}}, \theta) = \theta^{N_{\mathrm{H}}}(1-\theta)^{N_{\mathrm{T}}}$$

Also, the PDF for beta distributions is given as,

$$Beta(\theta, \alpha, \beta) = C.\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

The posterior is proportional to the product of likelihood and prior and putting their values we have,

$$P(\theta|\mathcal{D}) \propto P(\mathcal{D}|\theta)P(\theta) \tag{11}$$
$$\propto \theta^{N_{\mathrm{H}}+\alpha-1}(1-\theta)^{N_{\mathrm{T}}+\beta-1} \tag{12}$$
$$\propto Beta(\theta, N_{\mathrm{H}}+\alpha, N_{\mathrm{T}}+\beta) \tag{13}$$

$$\implies P(\theta|\mathcal{D}) \propto Beta(\theta, N_{\mathrm{H}}+\alpha, N_{\mathrm{T}}+\beta)$$

From the above coin example, the Beta distribution is a conjugate prior for Bernoulli or Binomial likelihoods.

Additional Note:
There are a large number of exponential families that serve as conjugate priors for different distributions due to their exponential mathematical form.

1. Beta distribution serves as conjugate prior for Bernoulli & Binomial likelihoods.
2. Normal distribution is a conjugate prior for itself, i.e. a Gaussian distribution is a conjugate prior for other Gaussian likelihoods, however, inverse gamma distribution also is a conjugate prior for Gaussian likelihood.
3. Dirichlet prior is used as conjugate for Multinomial distributions.