Recitation 2
Linear Regression and MLE/MAP

10-701: Introduction to Machine Learning (PhD)

09/12/2025

1 Linear Regression

In this section, we will consider the following linear regression model:

For each data point in $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$,

$$y_i = \boldsymbol{w}^T \boldsymbol{x}_i + \epsilon \text{ where } y_i, \epsilon \in \mathbb{R} \text{ and } \boldsymbol{w}, \boldsymbol{x}_i \in \mathbb{R}^{d+1}$$

In matrix notation, we can express this linear relationship for all data points as:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\epsilon} \text{ where } \boldsymbol{y}, \boldsymbol{\epsilon} \in \mathbb{R}^n, \boldsymbol{X} \in \mathbb{R}^{n \times (d+1)}, \text{ and } \boldsymbol{w} \in \mathbb{R}^{d+1}$$

1.1 Ordinary Least Squares (OLS)

In class, we saw that one way to optimize $\boldsymbol{w}$ is to minimize the least squares error:

$$\begin{aligned} \boldsymbol{w}_{\text{LS}}^* &= \arg\min_{\boldsymbol{w}} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||_2^2 \\ &= \arg\min_{\boldsymbol{w}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) \end{aligned}$$

1. Derive the least squares optimal solution $\boldsymbol{w}_{\text{LS}}^*$. You may assume any matrix inversion that naturally appears is possible.

2. Now let us consider the following: In general, when we have some matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{b} \in \mathbb{R}^n$, the orthogonal projection of $\boldsymbol{b}$ onto the column space of $\boldsymbol{A}$ can be done using the projection matrix $\boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T$. With this in mind, what can we say about $\boldsymbol{w}^*_{\text{LS}}$?

2 MLE/MAP

2.1 Definitions

- Likelihood: $\mathcal{L}(\theta) = \mathbb{P}(\mathcal{D}|\theta)$ and $l(\theta) = \log \mathbb{P}(\mathcal{D}|\theta)$
- Posterior: $\mathbb{P}(\theta|\mathcal{D}) = \frac{\mathbb{P}(\mathcal{D}|\theta)\mathbb{P}(\theta)}{\mathbb{P}(\mathcal{D})} \propto \mathbb{P}(\mathcal{D}|\theta)\mathbb{P}(\theta)$
- MLE estimate: $\theta_{MLE} = \arg\max_\theta \mathbb{P}(\mathcal{D}|\theta) = \arg\max_\theta \log \mathbb{P}(\mathcal{D}|\theta)$
- MAP estimate: $\theta_{MAP} = \arg\max_\theta \mathbb{P}(\mathcal{D}|\theta)\mathbb{P}(\theta) = \arg\max_\theta \log \mathbb{P}(\mathcal{D}|\theta) + \log \mathbb{P}(\theta)$

2.2 MLE/MAP Practice problem

Imagine you are a data scientist working in a hospital, and the emergency department is interested in knowing the probability of diagnosing any given patient that visits the ER with the flu. The following observed data is given to you (you know they are i.i.d.):

D_1	Monday	80/100 patients
D_2	Tuesday	15/200 patients
D_3	Wednesday	5/150 patients

Assume that the probability of k patients having the flu out of n total patients that visit the ER on any given day is determined by a binomial distribution. Namely

$$\mathbb{P}(patients_{flu} = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Using the data observed above, calculate p_{MLE}. Do you find this answer to be particularly useful or interesting? How is it similar to calculate p_{MLE} for a Bernoulli random variable?

Now we'll assume that we've entered flu season, and to incorporate this information to our estimate, we've decided to use a beta prior:

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

where $B(\alpha,\beta)$ is a normalizing constant. In particular, we'll use $\alpha = 101, \beta = 51$.

Find the posterior distribution and calculate the MAP estimate, p_{MAP}.

Now compare the MLE and MAP estimates, and interpret the meaning of Beta prior's parameters α and β in our context.

2.3 Gaussian MLE

Given that we have i.i.d samples $D = \{x_1, ..., x_N\}$, where each point is identically distributed according to a Gaussian distribution, find the MLE for the mean and variance. Hint: $p(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$