Recitation 2
Linear Regression and MLE/MAP

10-701: Introduction to Machine Learning (PhD)

02/02/2024

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.

We can solve by setting the derivative w.r.t $\boldsymbol{w}$ of the minimized statement to 0

$$\frac{\partial}{\partial \boldsymbol{w}}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}) = \frac{\partial}{\partial \boldsymbol{w}}(\boldsymbol{y}^T\boldsymbol{y}-(\boldsymbol{X}\boldsymbol{w})^T\boldsymbol{y}-\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{w})+(\boldsymbol{X}\boldsymbol{w})^T\boldsymbol{X}\boldsymbol{w})$$

In this situation, we can look at how to derive and/or combine the matrix products above. For $\boldsymbol{y}^T\boldsymbol{y}$ the derivative is just 0, due to it not involving w. Since $\boldsymbol{X}\boldsymbol{w}$ is size n x 1 and $\boldsymbol{y}$ is size n x 1, $(\boldsymbol{X}\boldsymbol{w})^T\boldsymbol{y}$ and $\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{w})$ are both scalars so they can just be combined.

$$\begin{aligned}
&= \frac{\partial}{\partial \boldsymbol{w}}(\boldsymbol{y}^T\boldsymbol{y}-(\boldsymbol{X}\boldsymbol{w})^T\boldsymbol{y}-\boldsymbol{y}^T(\boldsymbol{X}\boldsymbol{w})+(\boldsymbol{X}\boldsymbol{w})^T\boldsymbol{X}\boldsymbol{w}) \\
&= \frac{\partial}{\partial \boldsymbol{w}}(-2(\boldsymbol{X}\boldsymbol{w})^T\boldsymbol{y}+(\boldsymbol{X}\boldsymbol{w})^T\boldsymbol{X}\boldsymbol{w})
\end{aligned}$$

If we expand out $\boldsymbol{X}\boldsymbol{w}$, we get a n x 1 vector $[\boldsymbol{x_1}\boldsymbol{w}^T, \boldsymbol{x_2}\boldsymbol{w}^T, ..., \boldsymbol{x_n}\boldsymbol{w}^T]$ where x_i is the ith row of x.

Then, expanding $(\boldsymbol{X}\boldsymbol{w})^T\boldsymbol{X}\boldsymbol{w}$, we get the scalar $(\boldsymbol{x_1}\boldsymbol{w}^T)^2 + ... + (\boldsymbol{x_n}\boldsymbol{w}^T)^2$

Deriving by w for each i, we get $2(\boldsymbol{x_i}\boldsymbol{w}^T)x_i$. This scalar sum is equivalent to $2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w}$

Thus, we have the final derivative:

$$\begin{aligned}
&= \frac{\partial}{\partial \boldsymbol{w}}(-2(\boldsymbol{X}\boldsymbol{w})^T\boldsymbol{y}+(\boldsymbol{X}\boldsymbol{w})^T\boldsymbol{X}\boldsymbol{w}) \\
&= -2\boldsymbol{X}^T\boldsymbol{y}+2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w}
\end{aligned} \tag{1}$$

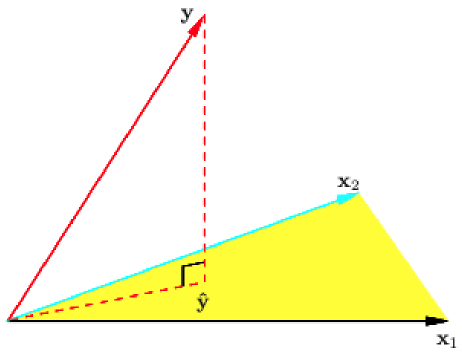
Solving for $-2\boldsymbol{X}^T\boldsymbol{y}+2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w}=0$, we get that $\boldsymbol{w}=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$

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 $\boldsymbol{w}^*_{\text{LS}}$?

Notice that since $\boldsymbol{w}^*_{\text{LS}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$, we have

$$\hat{\boldsymbol{y}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

This shows that $\hat{\boldsymbol{y}}$ is an orthogonal projection of $\boldsymbol{y}$ onto the column space of $\boldsymbol{X}$. That is, from a geometric perspective, we are finding the value of $\boldsymbol{w}$ that gives us this orthogonal projection matrix when minimizing the least squares error.


$\mathbf{y}$
$\mathbf{x}_2$
$\hat{\mathbf{y}}$
$\mathbf{x}_1$

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?

Define the log-likelihood function, since data is iid:

$$\begin{aligned}
\log \mathcal{L}(p) &= \log \mathbb{P}(\mathcal{D}|p) \\
&= \log \prod_{i=1}^{3} \mathbb{P}(\mathcal{D}|p) \\
&= \sum_{i=1}^{3} \log \mathbb{P}(\mathcal{D}|p) \\
&= \sum_{i=1}^{3} \log \left[\binom{n_i}{k_i} p^{k_i} (1-p)^{n_i - k_i} \right] \\
&= \sum_{i=1}^{3} \left[\log \binom{n_i}{k_i} + k_i \log p + (n_i - k_i) \log(1-p) \right] \\
&= \sum_{i=1}^{3} \log \binom{n_i}{k_i} + \log p \left[\sum_{i=1}^{3} k_i \right] + \log(1-p) \left[\sum_{i=1}^{3} (n_i - k_i) \right]
\end{aligned}$$

Take the partial derivative with respect to p:

$$\frac{\partial l(p)}{\partial p} = \frac{1}{p} \sum_{i=1}^{3} k_i - \frac{1}{1-p} \sum_{i=1}^{3} (n_i - k_i)$$

Set the partial derivative to zero and solve for p:

$$\frac{1}{p} \sum_{i=1}^{3} k_i - \frac{1}{1-p} \sum_{i=1}^{3} (n_i - k_i) = 0$$

$$p = \frac{\sum_{i=1}^{3} k_i}{\sum_{i=1}^{3} k_i + \sum_{i=1}^{3} (n_i - k_i)} = \frac{\sum_{i=1}^{3} k_i}{\sum_{i=1}^{3} n_i}$$

$$p_{MLE} = 100/450 \approx 0.22$$

The answer is the same as if you were to just intuit that the total number of patients with flu divided by the total number of patients that visited the ER. We didn't learn much, especially considering the disproportionate number of people that come to the ER with the flu on Monday. If we cared about making this prediction to better prepare for staffing or supplies needs, knowing the average percent of ER patients with flu isn't going to help us. The answer is the same as if we had observed 450 patients and modeled the outcome (flu) as a Bernoulli random variable.

$$p_{MLE} = 100/450 \approx 0.22$$

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}.

Let's first find the form of the posterior:

$$\begin{aligned}
\mathbb{P}(p \mid \mathcal{D}) &= \frac{\mathbb{P}(\mathcal{D} \mid p)\mathbb{P}(p)}{\mathbb{P}(\mathcal{D})} \\
&\propto \mathbb{P}(\mathcal{D} \mid p)\mathbb{P}(p) \\
&= \left(\prod_{i=1}^{3} \binom{n_i}{k_i} p^{k_i}(1-p)^{n_i-k_i}\right) \cdot \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1} \\
&= \left(\frac{1}{B(\alpha,\beta)} \prod_{i=1}^{3} \binom{n_i}{k_i}\right) \cdot p^{\alpha-1+\sum_{i=1}^{3} k_i}(1-p)^{\beta-1+\sum_{i=1}^{3}(n_i-k_i)} \\
&\propto p^{\alpha-1+\sum_{i=1}^{3} k_i}(1-p)^{\beta-1+\sum_{i=1}^{3}(n_i-k_i)}
\end{aligned}$$

It turns out that the beta distribution is the conjugate prior, so the posterior is another beta distribution, with $\alpha' = \alpha + \sum k_i$ and $\beta' = \beta + \sum(n_i - k_i)$.

Now let's calculate the MAP estimate by maximizing the posterior:

$$\begin{aligned}
p_{MAP} &= \arg\max_p \mathbb{P}(p \mid \mathcal{D}) \\
&= \arg\max_p \log \mathbb{P}(p \mid \mathcal{D}) \\
&= \arg\max_p \log \left(p^{\alpha-1+\sum_{i=1}^{3} k_i}(1-p)^{\beta-1+\sum_{i=1}^{3}(n_i-k_i)}\right) \\
&= \arg\max_p \quad \log p \cdot \left(\alpha - 1 \cdot + \sum_{i=1}^{3} k_i\right) + \log(1-p) \cdot \left(\beta - 1 + \sum_{i=1}^{3}(n_i - k_i)\right)
\end{aligned}$$

Taking the partial derivative with respect to p and setting equal to zero:

$$\begin{aligned}
\frac{1}{p}\left(\alpha - 1 \cdot + \sum_{i=1}^{3} k_i\right) - \frac{1}{1-p}\left(\beta - 1 + \sum_{i=1}^{3}(n_i - k_i)\right) &= 0 \\
p &= \frac{\alpha - 1 + \sum_{i=1}^{3} k_i}{(\alpha-1) + (\beta-1) + \sum_{i=1}^{3} n_i} \\
p_{MAP} &= \frac{(101-1) + 100}{(101-1) + (51-1) + 450} \approx 0.33
\end{aligned}$$

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

The answer to the MAP estimate is along same lines as the MLE estimate, but if we hypothetically had also previously observed $\alpha - 1$ flu patients and $\beta - 1$ non-flu patients. Since $\frac{\alpha-1}{\beta-1} = \frac{2}{3}$ was higher than our MLE estimate, our MAP estimate ends up moving in that direction and increases as well.

However, it's important to note that even if we keep the ratio of $\frac{\alpha-1}{\beta-1}$ the same, the larger we make α and β, the more the prior dominates the MAP estimate and ignores the observed data. Alternatively the smaller we make α and β, the more the MAP estimate ignores the prior and is dominated by the observed data.

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: $p(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Solution:

$$\begin{aligned}
\mu^* &= \arg\max_{\mu} \mathbb{P}(\mathcal{D}|\mu,\sigma) \\
&= \arg\max_{\mu} \prod_{i=1}^{n} \mathbb{P}(x_i|\mu,\sigma) \\
&= \arg\max_{\mu} \prod_{i=1}^{n} \left(\frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2} \right) \right) \\
&= \arg\max_{\mu} \left\{ \ln\left(\left(\frac{1}{\sigma\sqrt{2\pi}} \right)^n \right) + \sum_{i=1}^{n} -\frac{(x_i-\mu)^2}{2\sigma^2} \right\}
\end{aligned}$$

Then we can solve for μ^* by solve for $\partial/\partial\mu$ (since the function is concave)

$$\frac{\partial}{\partial\mu} \left\{ n\ln\left(\frac{1}{\sigma\sqrt{2\pi}} \right) + \sum_{i=1}^{n} -\left(\frac{(x_i-\mu)^2}{2\sigma^2} \right) \right\} = \frac{1}{2\sigma^2} \cdot 2 \sum_{i=1}^{n} (x_i - \mu)$$

Hence, we have μ^* be the solution of

$$\sum_{i=1}^{n} (x_i - \mu) = 0$$

That is, we have $\mu^* = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$.

For σ^*, we have something similar here

$$\begin{aligned} \sigma^* &= \arg\max_{\sigma} \mathbb{P}(\mathcal{D}|\mu, \sigma) \\ &= \cdots \text{ (same as above)} \\ &= \arg\max_{\sigma} \left\{ n \ln \left(\frac{1}{\sigma\sqrt{2\pi}} \right) + \sum_{i=1}^{n} \left(-\frac{(x_i - \mu)^2}{2\sigma^2} \right) \right\} \end{aligned}$$

We can solve σ^* by solving $\partial / \partial\sigma = 0$.

$$\frac{\partial}{\partial\sigma} \left\{ n \ln \left(\frac{1}{\sigma\sqrt{2\pi}} \right) + \sum_{i=1}^{n} \left(-\frac{(x_i - \mu)^2}{2\sigma^2} \right) \right\} = -\frac{n}{\sigma} + \left(\sum_{i=1}^{n} (x_i - \mu)^2 \right) \cdot \frac{1}{\sigma^3}$$

Hence, we have the solution of σ^* being the solution of

$$-\frac{n}{\sigma} + \left(\sum_{i=1}^{n} (x_i - \mu)^2 \right) \cdot \frac{1}{\sigma^3}$$

That is,

$$(\sigma^*)^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$$