Maximum Likelihood Estimation & Maximum A Posteriori Probability Estimation
Announcements

1. HW2 (Perceptron, PCA, K-means) will be out today
Recap on Perceptron

Binary classifier: \( \text{sign}(w^T x) \)

**The Perceptron Alg:**

**Initialize** \( w_0 = 0 \)

For \( t = 0 \to \infty \)

- feature \( x_t \) shows up
- We make a prediction \( \hat{y}_t = \text{sign}(w_t^T x_t) \)
- Check if \( \hat{y}_t \) equal to \( y_t \)
- We update \( w_{t+1} = w_t + 1(\hat{y}_t \neq y_t)y_tx_t \)

Q: how to apply this on a static dataset \( \mathcal{D} = \{x_i, y_i\}_{i=1}^n \)?

Q: If data has margin \( y_i(x_i^T w^*) \geq \gamma \), does it guarantee to converge to \( w^* \)?
Objective for today:

Understand the two common statistical learning framework: MLE and MAP
Outline for today:

1. Maximum Likelihood estimation (MLE)

2. Maximum a posteriori probability (MAP)
Ex 1: Estimating the probability of a coin flip

We toss a coin $n$ times (independently), we observe the following outcomes:

$$\mathcal{D} = \{y_i\}_{i=1}^n, y_i \in \{-1, 1\} \quad (y_i = 1 \text{ means head in } i\text{'s trial, -1 means tail})$$

Q: assume $y_i \sim \text{Bernoulli}(\theta^*)$, how to estimate $\theta^*$ given $\mathcal{D}$?

$$\hat{\theta} = \frac{\sum_{i=1}^n 1(y_i = 1)}{n}$$

Let's make this rigorous!
Maximum Likelihood Estimation

We toss a coin \( n \) times (independently), we observe the following outcomes:

\[ \mathcal{D} = \{y_i\}_{i=1}^{n}, y_i \in \{-1,1\} \quad (y_i = 1 \text{ means head in } i\text{'s trial, -1 means tail}) \]

If the probability of getting head is \( \theta \in [0,1] \), what is the probability of observing the data \( \mathcal{D} \) (i.e., likelihood)?

\[ P(\mathcal{D} | \theta) = \theta^{n_1}(1 - \theta)^{n-n_1} \]

**MLE Principle:** Find \( \theta \) that **maximizes the likelihood** of the data:

\[ \hat{\theta}_{mle} = \arg \max_{\theta \in [0,1]} P(\mathcal{D} | \theta) \]
Maximum Likelihood Estimation

We toss a coin n times (independently), we observe the following outcomes:

\[ \mathcal{D} = \{y_i\}_{i=1}^n, y_i \in \{-1, 1\} \quad (y_i = 1 \text{ means head in } i\text{'s trial, -1 means tail}) \]

**MLE Principle**: Find \( \theta \) that maximizes the likelihood of the data:

\[
\hat{\theta}_{mle} = \arg \max_{\theta \in [0,1]} P(\mathcal{D} | \theta) = \arg \max_{\theta \in [0,1]} \theta^{n_1}(1 - \theta)^{n - n_1}
\]

\[
= \arg \max_{\theta \in [0,1]} \ln(\theta^{n_1}(1 - \theta)^{n - n_1})
\]

\[
= \arg \max_{\theta \in [0,1]} n_1 \ln(\theta) + (n - n_1)\ln(1 - \theta) = \frac{n_1}{n}
\]
**Ex 2: Estimate the mean**

\[ \mathcal{D} = \{x_i\}_{i=1}^{n}, x_i \in \mathbb{R}^d \]

Assume data is from \( \mathcal{N}(\mu^*, I) \), want to estimate \( \mu^* \) from the data \( \mathcal{D} \).

Let’s apply the MLE Principle:

**Step 1:**

\[
P(\mathcal{D} | \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{(2\pi)^d}} \exp \left( -\frac{1}{2} (x_i - \mu)^\top (x_i - \mu) \right)
\]

**Step 2:** apply log and maximize the log-likelihood:

\[
\arg\max_{\mu} \sum_{i=1}^{n} - (x_i - \mu)^\top (x_i - \mu) \Rightarrow \hat{\mu}_{mle} = \sum_{i=1}^{n} x_i / n
\]
Q: Estimate the mean and variance

\[ \mathcal{D} = \{ x_i \}_{i=1}^n, x_i \in \mathbb{R} \]

Assume data is from \( \mathcal{N}(\mu^*, \sigma^2) \), want to estimate \( \mu^*, \sigma \) from the data \( \mathcal{D} \)

Let’s apply the MLE Principle:

Step 1: \( P(\mathcal{D} | \mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sigma \sqrt{2\pi}} \exp \left( -\frac{1}{2} \frac{(x_i - \mu)^2}{\sigma^2} \right) \)

Step 2: apply log and maximize the log-likelihood:

\[ \arg \max_{\mu, \sigma > 0} \sum_{i=1}^{n} \left( -\frac{(x_i - \mu)^2}{\sigma^2} - \ln(\sigma) \right) = ?? \]
Some properties of MLE

1. MLE is consistent: if our model assumption is correct (e.g., coin flip follows some Bernoulli distribution), then $\hat{\theta}_{mle} \rightarrow \theta^*$, as $n \rightarrow \infty$

2. When our model assumption is wrong (e.g., we use Gaussian to model data which is from some more complicated distribution), then MLE loses such guarantee
Outline for today:

1. Maximum Likelihood estimation (MLE)

2. Maximum a Posteriori Probability (MAP)
Ex: Estimating the probability of a coin flip

We toss a coin $n$ times (independently), we observe the following outcomes:

$$\mathcal{D} = \{y_i\}_{i=1}^{n}, y_i \in \{-1, 1\} \quad (y_i = 1 \text{ means head in } i\text{'s trial, } -1 \text{ means tail})$$

A Bayesian Statistician will treat the optimal parameter $\theta^*$ being a random variable:

$$\theta^* \sim P(\theta)$$

Example: $P(\theta)$ being a Beta distribution:

$$P(\theta) = \theta^{\alpha-1}(1 - \theta)^{\beta-1}/Z,$$

where $Z = \int_{\theta \in [0,1]} \theta^{\alpha-1}(1 - \theta)^{\beta-1} d\theta$
The Posterior distribution over $\theta$

Now, we have a prior $P(\theta)$, and we have a dataset $\mathcal{D} = \{y_i\}_{i=1}^n$, define posterior distribution:

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

Using Bayes rule, we get:

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

Posterior $\propto$ Prior $\times$ Likelihood
Maximum A Posteriori Probability estimation (MAP)

\[ P(\theta \mid \mathcal{D}) \propto P(\theta)P(\mathcal{D} \mid \theta) \]

\[ \hat{\theta}_{map} = \arg \max_{\theta \in [0,1]} P(\theta \mid \mathcal{D}) = \arg \max_{\theta \in [0,1]} P(\theta)P(\mathcal{D} \mid \theta) \]

\[ = \arg \max_{\theta \in [0,1]} \ln P(\theta) + \ln P(\mathcal{D} \mid \theta) \]
MAP for coin flip

\[ \hat{\theta}_{map} = \arg \max_{\theta \in [0,1]} \ln(P(\theta)P(\mathcal{D} | \theta)) \]

Step 1: specify Prior \( P(\theta) \propto \theta^{\alpha-1}(1 - \theta)^{\beta-1} \)

Step 2: data likelihood \( P(\mathcal{D} | \theta) = \theta^{n_1}(1 - \theta)^{n-n_1} \)

Step 3: Compute posterior \( P(\theta | \mathcal{D}) \propto \theta^{n_1+\alpha-1}(1 - \theta)^{n-n_1+\beta-1} \)

Step 4: Compute MAP \( \hat{\theta}_{map} = \frac{n_1 + \alpha - 1}{n + \alpha + \beta - 2} \)

\((\alpha - 1, \beta - 1)\) can be understood as some fictions flips: we had \(\alpha - 1\) hallucinated heads, and \(\beta - 1\) hallucinated tails
Some considerations on prior distributions

1. In coin flip example, when $n \to \infty$, $\hat{\theta}_{map} = \frac{n_1 + \alpha - 1}{n + \alpha + \beta - 2} \to \frac{n_1}{n}$ (i.e., $\hat{\theta}_{mle}$)

2. When $n$ is small and our prior is accurate, MAP can work better than MLE

3. In general, not so easy to set up a good prior....
Summary

\[ P(\theta | \mathcal{D}) \]

\[ P(\mathcal{D} | \theta) \]

\[ P(\theta) \]
Summary for today

1 MLE (frequentist perspective):

The ground truth $\theta^\star$ is unknown but fixed; we search for the parameter that makes the data as likely as possible

$$\arg\max_\theta P(\mathcal{D} | \theta)$$

2 MAP (Bayesian perspective):

The ground truth $\theta^\star$ treated as a random variable, i.e., $\theta^\star \sim P(\theta)$; we search for the parameter that maximizes the posterior

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