CS 340 Lec. 13: Maximum Likelihood

AD

February 2011

Problem

- You are given data $\left\{\mathbf{x}^i\right\}_{i=1}^{N}$ ($\left\{\mathbf{x}^i, y^i\right\}_{i=1}^{N}$ in the supervised learning case).
- You have a probabilistic model for the data; i.e. typically in most learning problem

$$p\left(\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^N \middle| \theta\right) = \prod_{i=1}^{N} p\left(\mathbf{x}^i \middle| \theta\right)$$

- **Aim**: you want to pick the best $\theta \in \Theta$.
- Two main approaches considered here: Maximum Likelihood and Bayesian.

Maximum Likelihood Parameter Estimation

- The most standard approach consists of selecting

$$\begin{aligned}\widehat{\theta}_{MLE} &= \arg\max_{\theta\in\Theta}\ p\left(\mathbf{x}^1,\mathbf{x}^2,...,\mathbf{x}^N\middle|\,\theta\right)\\ &= \arg\max_{\theta\in\Theta}\ \log\ p\left(\mathbf{x}^1,\mathbf{x}^2,...,\mathbf{x}^N\middle|\,\theta\right)\end{aligned}$$

- You select the value of $\theta\in\Theta$ that maximizes the probability of observing $\left(\mathbf{x}^1,\mathbf{x}^2,...,\mathbf{x}^N\right)$.
- **Example**: Assume independent $\left\{\mathbf{x}^i\right\}$ where $\mathbf{x}^i = x_1^i = x^i$ with

$$p\left(x\middle|\,\theta\right) = \theta^{\mathbb{I}(x=1)}\left(1-\theta\right)^{\mathbb{I}(x=0)}$$

then $\widehat{\theta}_{MLE} = \sum_{i=1}^N \mathbb{I}\left(x^i = 1\right)/N$.

Maximum Likelihood for Poisson Data

- **Example:** Assume you have independent Poisson observations $\{x^i\}_{i=1}^N$ such that

$$p(x|\theta) = e^{-\theta}\frac{\theta^x}{x!}$$

  for $\theta > 0$ and $x = 0, 1, 2, ...$
- In this case, we have

$$\begin{aligned} l(\theta) &= \log\ p\left(x^{1:N}\middle|\,\theta\right) \\ &= -N\theta + \log\theta\ \sum_{i=1}^N x^i - \sum_{i=1}^N \log x^i! \end{aligned}$$

- By setting $\frac{\partial l(\theta)}{\partial\theta} = 0$, we obtain

$$\widehat{\theta}_{MLE} = \frac{\sum_{i=1}^N x^i}{N}.$$

Maximum Likelihood for Gaussian Data

- **Example:** Assume you independent observations $\left\{x^i\right\}_{i=1}^N$ such that

$$p\left(x|\,\theta\right)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}\left(x-\mu\right)^2\right)$$

where $\theta=\left(\sigma^2,\mu\right)$.

- We have

$$\begin{aligned} l\left(\theta\right) &= \log\ p\left(x^{1:N}\middle|\,\theta\right) \\ &= -\frac{N}{2}\log\left(2\pi\sigma^2\right)-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(x^i-\mu\right)^2 \end{aligned}$$

- By setting $\frac{\partial l(\theta)}{\partial\mu}=0$ and $\frac{\partial l(\theta)}{\partial\sigma^2}=0$, we obtain

$$\widehat{\mu}_{MLE}=\frac{1}{N}\sum_{i=1}^{N}x^i,\ \ \widehat{\sigma^2}_{MLE}=\frac{1}{N}\sum_{i=1}^{N}\left(x^i-\widehat{\mu}_{MLE}\right)^2.$$

Maximum Likelihood for Multinomial Data

- **Example:** Assume you have independent observations $\left\{ \mathbf{x}^i \right\}_{i=1}^N$ such that

$$p\left(\mathbf{x} \middle| \theta \right) = \frac{P!}{\prod_{i=1}^{d} x_k!} \prod_{k=1}^{d} \theta_k^{x_k}$$

for $\theta_k > 0$, $\sum_{k=1}^{d} \theta_k = 1$ and $x_k = 0, 1, 2, ..., P$ with $\sum_k x_k = P$.

- In this case, we have

$$\begin{aligned} l\left(\theta\right) &= \log \ p\left(\mathbf{x}^{1:N} \middle| \theta \right) \\ &= \sum_{i=1}^{N} \log\left(\frac{P!}{\prod x_k^i!} \right) + \sum_{k=1}^{d} \left(\sum_{i=1}^{N} x_k^i \right) \log\theta_k \end{aligned}$$

- Be careful: It is a constrained optimization problem as $\sum_{k=1}^{d} \theta_k = 1$.

Maximum Likelihood for Multinomial Data

- We introduce a Lagrange multiplier λ and propose to maximize instead w.r.t θ and λ

$$l(\theta, \lambda) = l(\theta) + \lambda \left(1 - \sum_{k=1}^{d} \theta_k\right).$$

- Setting $\frac{\partial l(\theta,\lambda)}{\partial \lambda} = 0 \Rightarrow \sum_{k=1}^{d} \theta_k = 1$ and setting

$$\frac{\partial l(\theta, \lambda)}{\partial \theta_i} = 0 \Rightarrow \frac{\sum_{i=1}^{N} x_k^i}{\theta_k} - \lambda = 0 \Leftrightarrow \lambda \theta_k = \sum_{i=1}^{N} x_k^i$$

- It follows that, as $\sum_{k=1}^{d} \theta_k = 1$, then $\lambda = \left(\sum_{k=1}^{d} \sum_{i=1}^{N} x_k^i\right)$

$$\widehat{\theta}_{k,MLE} = \frac{\sum_{i=1}^{N} x_k^i}{\lambda} = \frac{\sum_{i=1}^{N} x_k^i}{\sum_{i=1}^{N} \sum_{k=1}^{d} x_k^i}.$$

Application to Naive Bayes

- We assume that we have N data $\left(\mathbf{x}^i, y^i\right) \in \mathbb{N}^d \times \{0,1\}^C$ and we use the model
$$p\left(\mathbf{x}, y=c \middle| \theta\right) = \pi_c \frac{P!}{\prod_{i=1}^{d} x_k!} \prod_{k=1}^{d} \theta_{k,c}^{x_k}$$
where $\theta = \left(\pi_1, ..., \pi_C, \theta_{1,1}, ..., \theta_{d,1}, \cdots, \theta_{1,C}, ..., \theta_{d,C}\right)$.
- We have
$$\begin{aligned} l\left(\theta\right) &= \textstyle\sum_{k=1}^{n} \log p\left(\mathbf{x}^i, y^i \middle| \theta\right) \\ &= \textstyle\sum_{c=1}^{C} \left(\sum_{k=1}^{N} \mathbb{I}\left(y_k^i = c\right)\right) \left\{\log \pi_c + \sum_{k=1}^{d} x_k^i \log \theta_{k,c} + Cste\right\} \end{aligned}$$
yields with $N_c = \sum_{k=1}^{N} \mathbb{I}\left(y_k^i = c\right)$
$$\widehat{\pi}_{c,MLE} = \frac{N_c}{N}, \quad \widehat{\theta}_{k,c,MLE} = \frac{\sum_{i=1}^{N} x_k^i \mathbb{I}\left(y_k^i = c\right)}{\sum_{k=1}^{d} \sum_{i=1}^{N} x_k^i \mathbb{I}\left(y_k^i = c\right)}.$$

Asymptotics of Maximum Likelihood Estimate

- Assume you have independent data $\left\{\mathbf{x}^i\right\}_{i=1}^N$ distributed according to $p\left(\mathbf{x}|\,\theta^*\right)$; i.e. θ^* is the true parameter. Under regularity assumptions, we have $\widehat{\theta}_{MLE} \rightarrow \theta^*$ as $N \rightarrow \infty$.
- This follows from the fact that first

$$\frac{l\left(\theta\right)}{N} = \frac{1}{N}\sum_{i=1}^{N} \log p\left(\mathbf{x}^i \middle|\, \theta\right) \underset{N\rightarrow\infty}{\rightarrow} \bar{l}\left(\theta\right) = \int p\left(\mathbf{x}|\,\theta^*\right) \log p\left(\mathbf{x}|\,\theta\right) d\mathbf{x}.$$

- Second, the average log-likelihood $\bar{l}\left(\theta\right)$ is maximized θ^*; for any $\theta \in \Theta$ as

$$\begin{aligned} \bar{l}\left(\theta\right) - \bar{l}\left(\theta^*\right) &= \int p\left(\mathbf{x}|\,\theta^*\right) \log \frac{p\left(\mathbf{x}|\,\theta\right)}{p\left(\mathbf{x}|\,\theta^*\right)} d\mathbf{x} \\ &\leq \log \left(\int p\left(\mathbf{x}|\,\theta^*\right) \frac{p\left(\mathbf{x}|\,\theta\right)}{p\left(\mathbf{x}|\,\theta^*\right)} d\mathbf{x} \right) \quad \text{(Jensen's inequality)} \\ &\leq 0. \end{aligned}$$

Limitations of Maximum Likelihood

- Maximum likelihood estimation overfits!
- Suppose we are N Bernoulli data such that $\widehat{\theta}_{MLE} = \sum_{i=1}^{N} x^i / N = 0$ then we have
$$p\left(x = 1 \middle| \widehat{\theta}_{MLE}\right) = 0.$$
- Similarly, suppose we have N multinomial data such that $\widehat{\theta}_{k,MLE} = \sum_{i=1}^{N} x_k^i / \sum_{i=1}^{N} \sum_{k=1}^{d} x_k^i = 0$ then we have
$$p\left(x_1, ..., x_{k-1}, x_k = 1, x_{k+1}, ..., x_d \middle| \widehat{\theta}_{MLE}\right) = 0.$$
- Hence if we have not observed such events in our training set, we predict that we will never observed them, ever!
- Failing to predict that certain events are possible is analogous to a problem in philosophy called the black swan paradox. This is based on the ancient Western conception that all swans were white. In that context, a black swan was a metaphor for something that could not exist. (Black swans were discovered in Australia by European explorers in the 17th Century.)