Parametric Models: from data to models

Pradeep Ravikumar (Instructor), HMW-Alexander (Noter)

January 24, 2017

Back to Index

Contents

Resources

- Lecture

---

1 Recall Model-based ML

Model-based ML

2 Model Learning: Data to Model

Questiongs:

- What are the principles in going from data to model?
- What are the guarantees of these methods?

2.1 Bernoulli Distribution Example

- Bernoulli distribution model
  - X is a random variable with Bernoulli distribution when:
    - X takes values in $\{0, 1\}$
    - $P(X = 1) = \theta,\ P(X = 0) = 1 - \theta$
    - Where $\theta \in [0, 1]$
- Draw **independent** samples that are **identically distributed** from same distribution model, Bernoulli distribution.
  - If we observe an event $X \in \{0, 1\}$, its probability $P(X)$ is $\theta^X(1-\theta)^{1-X}$
  - Then the probability of data:

$$\begin{aligned} \mathbb{P}(X_1, X_2, ..., X_n; \theta) &= \prod_{i=1}^n P(X_i) \\ &= \prod_{i=1}^n p^{X_i}(1-p)^{1-X_i} \\ &= p^{\sum_{i=1}^n X_i}(1-p)^{n-\sum_{i=1}^n X_i} \\ &= p^{n_1}(1-p)^{n-n_1} \end{aligned} \tag{1}$$

- Maximum Likelihood ($p(D|\theta)$) Estimator (MLE)
  - Choose θ that maximizes the probability of observed data.

$$\begin{aligned} \hat{\theta} &= \arg\max_\theta \mathbb{P}(X_1, \ldots, X_n; \theta) \\ &= \arg\max_\theta \theta^{n_1}(1-\theta)^{n-n_1} \\ &= \arg\max_\theta n_1 \log\theta + (n - n_1)\log(1-\theta) \\ &\Rightarrow \frac{n_1}{\hat{\theta}} - \frac{n-n_1}{1-\hat{\theta}} = 0 \\ &\Rightarrow \hat{\theta}_{MLE} = \frac{n_1}{n} \end{aligned} \tag{2}$$

  - MLE for parametric models
    - Data: $X_1, X_2, \ldots, X_n$
    - Model: $P(X|\theta)$ with parameters θ
    - Assumption: data drawn **i.i.d** from distribution $P(X|\theta^*)$ for some unknown θ^*
    - Mission: recover θ^* from data $X_1, X_2, \ldots, X_n$
    - Likelihood function: $L(\theta) := \prod_{i=1}^n P(X_i|\theta)$
    - Maximum Likelihood Estimator (MLE): find that parameter θ that would maximize the likelihood of θ.

2.2 How good is this MLE?

- Consistency:
  - As we sample more and more times, we want our estimator to converge (in probability) to the true probability.
  - For Bernoulli distribution example, we get the $\hat{\theta} = \frac{1}{n}\sum_{i=1}^n X_i \to \theta$ in probability as $n \to \infty$ by the **Law of Large Numbers**[1].
  - An estimator $\hat{\theta}(X_1, \ldots, X_n)$ where $X_i \sim P(X; \theta^*)$ is consistent if $\hat{\theta} \to \theta^*$ in probability as $n \to \infty$.
- Unbiasedness:
  - The estimator $\hat{\theta}$ is random: it depends on the samples drawn from a random distribution model with parameter θ. It would be great if the expectation $\mathbb{E}[\hat{\theta}]$ of the estimator $\hat{\theta}$ be equal to the "true" probability. This property is called unbiasedness.

[1] It does not apply to distributions for whom Expected values do not exist. One example of such a distribution is the Cauchy distribution where the mean and the variance are undefined.

– For Bernoulli example:

$$\begin{aligned}\mathbb{E}(\hat{\theta}) &= \mathbb{E}(\tfrac{n_1}{n}) \\ &= \mathbb{E}(\tfrac{\sum_{i=1}^{n} X_i}{n}) \\ &= \tfrac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) \\ &= \mathbb{E}(X_1) \\ &= \theta\end{aligned} \tag{3}$$

2.3 Gaussian Distribution Example

Gaussian Distribution:

$$P(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\exp(-\frac{(x-\mu)^2}{2\sigma^2}) = \mathcal{N}(\mu,\sigma^2)$$

- Affine transformation:
  - $X \sim \mathcal{N}(\mu,\sigma^2)$
  - $Y = aX + b \sim \mathcal{N}(a\mu+b, a^2\sigma^2)$
- Sum of Gaussians:
  - $X \sim \mathcal{N}(\mu_X,\sigma_X^2)$, $Y \sim \mathcal{N}(\mu_Y,\sigma_Y^2)$
  - $Z = X + Y \sim \mathcal{N}(\mu_X+\mu_Y, \sigma_X^2+\sigma_Y^2)$

MLE for Gaussian mean and variance:

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

2.3.1 The Biased Variance of a Gaussian

The unbiased variance estimator: $\hat{\sigma}^2_{unbiased} = \frac{n}{n-1}\hat{\sigma}^2_{MLE}$

Proof:

$$\begin{aligned}\mathbb{E}(\sigma^2_{MLE}) &= \mathbb{E}(\tfrac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2) \\ &= \tfrac{1}{n}\mathbb{E}(\sum_{i=1}^{n}(x_i^2 - 2x_i\hat{\mu}+\hat{\mu}^2)) \\ &= \tfrac{1}{n}\mathbb{E}(\sum_{i=1}^{n}x_i^2 - \sum_{i=1}^{n}2x_i\hat{\mu} + \sum_{i=1}^{n}\hat{\mu}^2) \\ &= \tfrac{1}{n}\mathbb{E}(\sum_{i=1}^{n}x_i^2 - 2n\hat{\mu}^2 + n\hat{\mu}^2) \\ &= \tfrac{1}{n}\mathbb{E}(\sum_{i=1}^{n}x_i^2 - n\hat{\mu}^2) \\ &= \tfrac{1}{n}\sum_{i=1}^{n}\mathbb{E}(x_i^2) - \mathbb{E}(\hat{\mu}^2) \\ &= \mathbb{E}(x_i^2) - \mathbb{E}(\hat{\mu}^2) \\ &= (\sigma^2(x_i)+\mathbb{E}(x_i)^2) - (\sigma^2(\hat{\mu})+\mathbb{E}(\hat{\mu})^2) \\ &= \sigma^2(x_i) - \sigma^2(\hat{\mu}) \\ &= \sigma^2(x_i) - \sigma^2(\tfrac{1}{n}\sum_{i=1}^{n}x_i) \\ &= \sigma^2(x_i) - \tfrac{1}{n^2}\sigma^2(\sum_{i=1}^{n}x_i) \\ &= \sigma^2(x_i) - \tfrac{1}{n^2}n\sigma^2(x_i) \\ &= \tfrac{n-1}{n}\sigma^2(x_i)\end{aligned} \tag{4}$$

3 Convergence Rates of Estimator

3.1 Simple Bound (Hoeffding's Inequality)

In probability theory, Hoeffding's inequality provides an upper bound on the probability that the sum of random variables deviates from its expected value. It can be applied to the important special case of identically distributed Bernoulli random variables.

- Let $X_1,\ldots,X_n$ be independent random variables bounded by the interval $[0,1] : 0 \le X_i \le 1$.
- and $\hat{\theta} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- then $\forall\epsilon > 0,\ P(|\hat{\theta} - \mathbb{E}(\hat{\theta})| \ge \tau) \le \exp(-2n\epsilon^2)$

3.2 PAC (Probably Approximate Correct) Learning

PAC is a learning framework. Its initials stand for: Probably Approximately Correct. PAC learning aims to provide bounds (worst case estimates) on the size of the dataset.

The terminology 'Probably Approximately Correct' comes from the requirement that with high probability (greater than 1-delta) the error rate(epsilon) will be small.

PAC bounds are very conservative, i.e they strongly over-estimate the size of the dataset required to give good generalization.

A more detailed version about PAC theory include origin, introduction, framework, etc. can be found at: http://web.cs.iastate.edu/~honavar/pac.pdf

Besides, one very interesting comment I have seen is that "PAC is the bridge between statistic and machine learning."

4 Computational Issues of MLE

When number of parameters, or number of samples n is large, computing the MLE is a large-scale optimization problem.