Probabilistic models

-

Maximum Likelihood Estimation

Marylou Gabrié

CDS, NYU

March 9, 2021

The Data: Assumptions So Far in this Course

- Our usual setup is that (x, y) pairs are drawn **i.i.d. from** $\mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$.
- So far ridge/lasso/ regression, optimization, SVMs, and kernel methods are applicable for arbitrary training data sets $\mathcal{D}:(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X}\times\mathcal{Y}$.
  - i.e. $\mathcal{D}$ could be created by hand, by an adversary, or randomly.
- How have we used this assumption so far?
  - motivates empirical risk minimization
  - ties test performance to performance on new data when deployed
- We rely on the i.i.d. $\mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$ assumption when it comes to **generalization** only.

Probabilistic Models: Use Assumptions on the Data for Learning

- Observations y are drawn i.i.d. from a distribution $\mathcal{P}_{\mathcal{Y}}$
  $\rightarrow$ **Maximum likelihood estimation** (First topic of week 6)
- Model how y depends on x
  $\rightarrow$ **Conditional probability models** $p(y|x)$ (Second topic of week 6)
- Incorporate prior knowledge and estimate uncertainty on the prediction
  $\rightarrow$ **Bayesian approaches** (Topic of week 7)

Maximum Likelihood Estimation: Contents

Table of Contents

Estimating a Probability Distribution: Setting

For the moment we only assume that we have one variable y.

- Let $p(y)$ represent a probability distribution on $\mathcal{Y}$.
- $p(y)$ is **unknown** and we want to **estimate** it.
- Assume that $p(y)$ is either a
  - probability density function on a continuous space $\mathcal{Y}$, or a
  - probability mass function on a discrete space $\mathcal{Y}$.
- Typical $\mathcal{Y}$'s:
  - $\mathcal{Y} = \mathbf{R}$; $\mathcal{Y} = \mathbf{R}^d$ [typical continuous distributions]
  - $\mathcal{Y} = \{-1, 1\}$ [e.g. binary classification]
  - $\mathcal{Y} = \{0, 1, 2, \ldots, K\}$ [e.g. multiclass problem]
  - $\mathcal{Y} = \{0, 1, 2, 3, 4 \ldots\}$ [unbounded counts]

Evaluating a Probability Distribution Estimate

- Before we talk about estimation, let's talk about evaluation.
- Somebody gives us an estimate of the probability distribution

$$\hat{p}(y).$$

- How can we evaluate how good it is?
- We want $\hat{p}(y)$ to be descriptive of **future** data.

Likelihood of a Predicted Distribution

- Suppose we have

$$\mathcal{D} = (y_1, \ldots, y_n) \text{ sampled i.i.d. from true distribution } p(y).$$

- Then the **likelihood** of $\hat{p}$ for the data $\mathcal{D}$ is defined to be

$$\hat{p}(\mathcal{D}) = \prod_{i=1}^{n} \hat{p}(y_i).$$

The probability of observing $\mathcal{D}$ under the estimate $\hat{p}$.

How are we going to construct an estimate of $\hat{p}(y)$?

Table of Contents

Parametric Models

Definition

A **parametric model** is a set of probability distributions indexed by a parameter $\theta \in \Theta$. We denote this as

$$\{p(y;\theta) \mid \theta \in \Theta\},$$

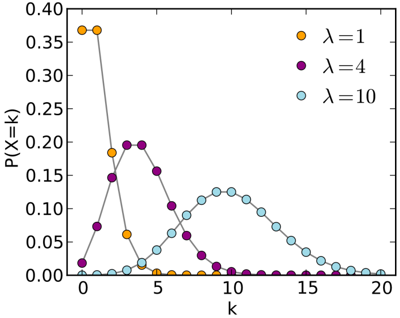
where θ is the **parameter** and Θ is the **parameter space**.

- Below we'll give some examples of common parametric models.
  - But it's worth doing research to find a parametric model most appropriate for your data.
- We'll sometimes say **family of distributions** for a probability model.

Poisson Family

- Support $\mathcal{Y} = \{0, 1, 2, 3, \ldots\}$.
- Parameter space: $\{\lambda \in \mathbf{R} \mid \lambda > 0\}$
- Probability mass function on $k \in \mathcal{Y}$:

$$p(k;\lambda) = \lambda^k e^{-\lambda} / (k!)$$



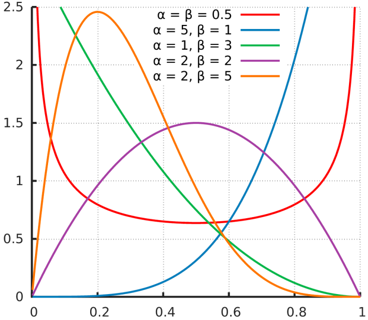
- Examples: Number of random i.i.d. events in a given time/over an interval
  - Radioactive decay of atoms over a year
  - Number of taxi cab pickups at Penn Station in an evening

Figure is "Poisson pmf" by Skbkekas - Own work. Licensed under CC BY 3.0 via Wikimedia Commons - http://commons.wikimedia.org/wiki/File:Poisson_pmf.svg#/media/File:Poisson_pmf.svg.

Beta Family

- Support $\mathcal{Y} = (0, 1)$. [The unit interval.]
- Parameter space: $\{\theta = (\alpha, \beta) \mid \alpha, \beta > 0\}$
- Probability density function on $y \in \mathcal{Y}$:

$$p(y; a, b) = \frac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha, \beta)}$$



- Examples: Spending of a resource over a interval.
  - Project management

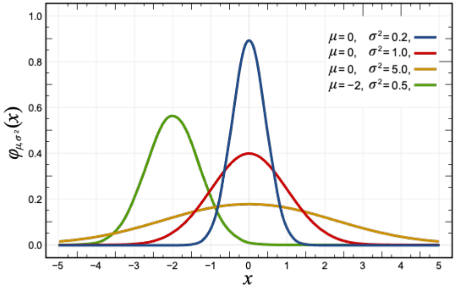
Figure by Horas based on the work of Krishnavedala (Own work) [Public domain], via Wikimedia Commons.

Gaussian Family

- Support $\mathcal{Y} \in \mathbf{R}$.
- Parameter space: $\{\theta = (\mu, \sigma^2) \mid \mu \in \mathbf{R}, \sigma^2 > 0\}$
- Probability density function on $y \in \mathcal{Y}$:

$$p(y; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(y-\sigma)^2/2\sigma^2}.$$

- Also named "normal" distribution, noted $\mathcal{N}(\mu, \sigma^2)$



- Examples: sum of i.i.d random variables (Central limit theorem)
  - Cumulated gain from random independent coin flips

Figure from Wikipedia https://en.wikipedia.org/wiki/Gaussian_function.

Multivariate Distributions

- Above we only cited examples of univariate distributions
- Sometimes we need multivariate distributions $p(y;\theta)$ for $y = (y_1, \cdots, y_d) \in \mathbf{R}^d$:
  - If y_is are independent $p(y;\theta) = \prod_{i=1}^{d} p(y_i;\theta_i)$
  - If there are correlations, we have to treat the problem in dimension d.

- Example:
  Multivariate Gaussian Distribution
  - In 2d: $y \in \mathbf{R}^2$, $p(y;\theta) = \mathcal{N}(\mu;\Sigma)$
  - Parameters:
    - Mean vector $\mu \in \mathbf{R}^2$
    - Covariance matrix $\Sigma \in \mathbf{R}^{2\times 2}$

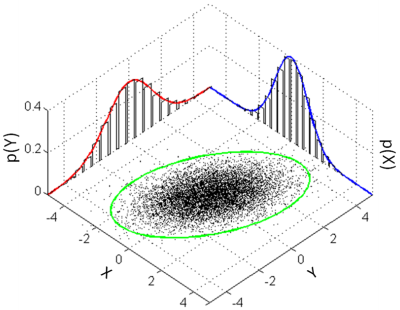


Figure from Wikipedia https://en.wikipedia.org/wiki/Gaussian_function.

Table of Contents

Likelihood in a Parametric Model

Suppose we have a parametric model $\{p(y;\theta) \mid \theta \in \Theta\}$ and a sample $\mathcal{D} = (y_1, \ldots, y_n)$.

- The **likelihood** of parameter estimate $\hat{\theta} \in \Theta$ for sample $\mathcal{D}$ is

$$p(\mathcal{D};\hat{\theta}) = \prod_{i=1}^{n} p(y_i;\hat{\theta}).$$

- In practice, we prefer to work with the **log-likelihood**. Same maximizer, but

$$\log p(\mathcal{D};\hat{\theta}) = \sum_{i=1}^{n} \log p(y_i;\hat{\theta}),$$

  and sums are easier to work with than products.

Maximum Likelihood Estimation

- Suppose $\mathcal{D} = (y_1, \ldots, y_n)$ is an i.i.d. sample from some distribution.

Definition

A **maximum likelihood estimator (MLE)** for θ in the model $\{p(y;\theta) \mid \theta \in \Theta\}$ is

$$\begin{aligned}\hat{\theta} &\in \underset{\theta\in\Theta}{\arg\max} \log p(\mathcal{D}, \hat{\theta}) \\ &= \underset{\theta\in\Theta}{\arg\max} \sum_{i=1}^{n} \log p(y_i;\theta).\end{aligned}$$

Maximum Likelihood Estimation

- Finding the MLE is an **optimization problem**.
- For some model families, calculus gives a closed form for the MLE.
- Can also use numerical methods we know (e.g. SGD).

MLE Existence

- In certain situations, the MLE may not exist.
- But there is usually a good reason for this.
- e.g. Gaussian family $\{\mathcal{N}(\mu, \sigma^2) \mid \mu \in \mathbf{R}, \sigma^2 > 0\}$
- We have a single observation y.
- Is there an MLE?
- Taking $\mu = y$ and $\sigma^2 \to 0$ drives likelihood to infinity.
- MLE doesn't exist.

Example: MLE for Poisson

- Observed counts $\mathcal{D} = (k_1, \ldots, k_n)$ for taxi cab pickups over n weeks.
    - k_i is number of pickups at Penn Station Mon, 7-8pm, for week i.
- We want to fit a Poisson distribution to this data.
- The Poisson log-likelihood for a single count is

$$\begin{aligned} \log\left[p(k;\lambda)\right] &= \log\left[\frac{\lambda^k e^{-\lambda}}{k!}\right] \\ &= k\log\lambda - \lambda - \log\left(k!\right) \end{aligned}$$

- The full log-likelihood is

$$\log p(\mathcal{D},\lambda) = \sum_{i=1}^{n}\left[k_i\log\lambda - \lambda - \log\left(k_i!\right)\right].$$

Example: MLE for Poisson

- The full log-likelihood is

$$\log p(\mathcal{D},\lambda) \quad = \quad \sum_{i=1}^{n}\left[k_i \log\lambda - \lambda - \log\left(k_i!\right)\right]$$

- First order condition gives

$$\begin{aligned} 0 = \frac{\partial}{\partial\lambda}\left[\log p(\mathcal{D},\lambda)\right] \quad &= \quad \sum_{i=1}^{n}\left[\frac{k_i}{\lambda} - 1\right] \\ \Longrightarrow \lambda \quad &= \quad \frac{1}{n}\sum_{i=1}^{n} k_i \end{aligned}$$

- So MLE $\hat{\lambda}$ is just the mean of the counts.

Estimating Distributions, Overfitting, and Hypothesis Spaces

- Just as in classification and regression, MLE can overfit!
- Example Probability Models: Penn Station, Mon-Fri 7-8pm
  - $\mathcal{F} = \{\text{Poisson distributions}\}$.
  - $\mathcal{F} = \{\text{Negative binomial distributions}\}$.
- How to judge which model works the best?
- Choose the model with the **highest likelihood on validation set**.
  - Test Set Log Likelihood for Penn Station, Mon-Fri 7-8pm

Method	Test Log-Likelihood
Poisson	-392.16
Negative Binomial	-188.67