Conditional Probability Models

David Rosenberg

New York University

March 30, 2015

Estimating a Probability Distribution: Setting

- 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 classificaiton]
  - $\mathcal{Y} = \{0, 1, 2, \ldots, K\}$ [e.g. multiclass problem]
  - $\mathcal{Y} = \{0, 1, 2, 3, 4 \ldots\}$ [unbounded counts]

Evaluating a Probabilty 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

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

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

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

- We'll write this as

$$L_{\mathcal{D}}(\hat{p}) := \hat{p}(\mathcal{D})$$

- Special case: If $\hat{p}$ is a probability mass function, then
  - $L_{\mathcal{D}}(\hat{p})$ is the probability of $\mathcal{D}$ under $\hat{p}$.

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**.

- Sometimes people began their analysis with something like:

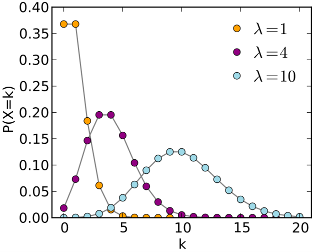
  Suppose the data are generated by a distribution in parametric family $\mathcal{F}$ (e.g. a Poisson family).

- Our perspective is different, at least conceptually:
  - We don't make any assumptions about the data generating distribution.
  - We use a parametric model as a **hypothesis space**.
  - (More on this later.)

Poisson Family

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

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



Beta Family

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

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

Figure by Horas based on the work of Krishnavedala (Own work) [Public domain], via Wikimedia Commons`http://taps-graph-review.wikispaces.com/Box+and+Whisker+Plots`.

Gamma Family

- Support $\mathcal{Y} = (0, \infty)$. [Positive real numbers]
- Parameter space: $\{\theta = (k, \theta) \mid k > 0, \theta > 0\}$
- Probability density function on $y \in \mathcal{Y}$:

$$p(y; k, \theta) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-x/\theta}.$$

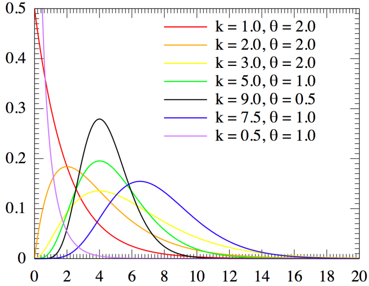


Figure from Wikipedia.

Maximum Likelihood Estimation

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

Definition

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

$$\hat{\theta} = \underset{\theta \in \Theta}{\arg\max}\, L_{\mathcal{D}}(\theta) = \underset{\theta \in \Theta}{\arg\max} \prod_{i=1}^{n} p(y_i, \theta).$$

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

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

and sums are easier to work with than products.

Maximum Likelihood Estimation

- Finding the MLE is an optimization problem.
- For some model families, calculus gives closed form for MLE.
- Otherwise, we can use the numerical methods we know (e.g. SGD).
- Note: In certain situations, the MLE may not exist.
  - But there is usually a good reason for this.
- e.g. Gaussian family $\left\{\mathcal{N}(\mu, \sigma^2 \mid \mu \in \mathbf{R}, \sigma^2 > 0\right\}$, Single observation y.
  - Take $\mu = y$ and $\sigma^2 \to 0$ drives likelihood to infinity. MLE doesn't exist.

Example: MLE for Poisson

- Suppose we've observed some counts $\mathcal{D} = \{k_1, \ldots, k_n\} \in \{0, 1, 2, 3, \ldots\}$.
- The Poisson log-likelihood for a single count is

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

- The full log-likelihood is

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

Example: MLE for Poisson

- The full log-likelihood is

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

- First order condition gives

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

$$\implies \lambda = \frac{1}{n} \sum_{i=1}^{n} k_i$$

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

Test Set Log Likelihood for Penn Station, Mon-Fri 7-8pm

Method	Test Log-Likelihood
Poisson	-392.16
Negative Binomial	-188.67
Histogram (Bin width $=7$)	$-\infty$
95% Histogram +.05 NegBin	-203.89

Probability Estimation as Statistical Learning

- Output space $\mathcal{Y}$ (containing observations from distribution P)
- **Action space**
  $\mathcal{A} = \{p(y) \mid p$ is a probability density or mass function on $\mathcal{Y}\}$.
- How to encode our objective of "high likelihood" as a loss function?
- Define loss function as the negative log-likelihood of y under $p(\cdot)$:

$$\begin{aligned} \ell : \quad \mathcal{A} \times \mathcal{Y} & \to \mathbf{R} \\ (p, y) & \mapsto -\log p(y) \end{aligned}$$

Probability Estimation as Statistical Learning

- The risk of p is

$$R(p) = \mathbb{E}_Y \left[-\log p(Y) \right].$$

- The empirical risk of p for a sample $\mathcal{D} = \{y_1, \ldots, y_n\} \in \mathcal{Y}$ is

$$\hat{R}(p) = -\sum_{i=1}^{n} \log p(y_i),$$

  which is exactly the **log-likelihood of** p **for the data** $\mathcal{D}$.

- Therefore, MLE is just an empirical risk minimizer!

Estimation Distributions, Overfitting, and Hypothesis Spaces

- Just as in classification and regression, MLE (i.e. ERM) can overfit!
- Example Hypothesis Spaces / Probability Models:
  - $\mathcal{F} = \{$Poisson distributions$\}$.
  - $\mathcal{F} = \{$Negative binomial distributions$\}$.
  - $\mathcal{F} =\{$Histogram with arbitrarily many bins$\}$ [will likely overfit for continuous data]
  - $\mathcal{F} =\{$Histogram with 10 bins$\}$
  - $\mathcal{F} = \{$Depth 5 decision trees with histogram estimates in leaves$\}$
- How to judge with hypothesis space works the best?
- Choose the model with the highest likelihood for a test set.

Generalized Regression / Conditional Distribution Estimation

- Given X, predict *probability distribution* $p(Y \mid X = x)$
- How do we represent the probability distribution?
- We'll consider *parametric families* of distributions.
  - distribution represented by parameter vector
- Examples:
  1. Logistic regression (Bernoulli distribution)
  2. Probit regression (Bernoulli distribution)
  3. Poisson regression (Poisson distribution)
  4. Linear regression (Normal distribution, fixed variance)
  5. Generalized Linear Models (GLM) (encompasses all of the above)
  6. Generalized Additive Models (GAM)
  7. Generalized Boosting Models (GBM)

Generalized Regression as Statistical Learning

- Input space $\mathcal{X}$
- Output space $\mathcal{Y}$
- All pairs (X, Y) are independent with distribution $P_{\mathcal{X}\times\mathcal{Y}}$.
- **Action space** $\mathcal{A} = \{p(y) \mid p$ is a probability density or mass function on $\mathcal{Y}\}$.
- Hypothesis spaces comprise decision functions $f : \mathcal{X} \to \mathcal{A}$.
  - Given an $x \in \mathcal{X}$, predict a probability distribution $p(y)$ on $\mathcal{Y}$.
- Loss function as before:

$$\begin{aligned} \ell : \quad \mathcal{A} \times \mathcal{Y} &\to \mathbf{R} \\ (p, y) &\mapsto -\log p(y) \end{aligned}$$

Generalized Regression as Statistical Learning

- The risk of decision function $f:\mathcal{X}\to\mathcal{A}$

$$R(f)=-\mathbb{E}_{X,Y}\log\left[f(X)\right](Y),$$

where $f(X)$ is a PDF or PMF on $\mathcal{Y}$, and we're evaluating it on Y.

- The empirical risk of f for a sample $\mathcal{D}=\{y_1,\ldots,y_n\}\in\mathcal{Y}$ is

$$\hat{R}(f)=-\sum_{i=1}^{n}\log\left[f(x_i)\right](y_i).$$

This is called the negative **conditional log-likelihood**.