

Conditional Probability Models

David Rosenberg

New York University

April 5, 2017

Maximum Likelihood Recap

Maximum Likelihood Estimation

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

Definition

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

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

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

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

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.
- Can also use 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 $\{\mathcal{N}(\mu, \sigma^2 \mid \mu \in \mathbf{R}, \sigma^2 > 0)\}$, Single observation y .
 - Take $\mu = y$ and $\sigma^2 \rightarrow 0$ drives likelihood to infinity. MLE doesn't exist.

Bernoulli Regression

Probabilistic Binary Classifiers

- Setting: $\mathcal{X} = \mathbf{R}^d$, $\mathcal{Y} = \{0, 1\}$
- For each x , need to predict a distribution on $\mathcal{Y} = \{0, 1\}$.
- What kind of parametric distribution could be supported on $\{0, 1\}$?
- Not a lot of choices....
- Bernoulli!
- For each x ,
 - predict the Bernoulli parameter $\theta = p(y = 1 \mid x)$.

Linear Probabilistic Classifiers

- Setting: $\mathcal{X} = \mathbf{R}^d$, $\mathcal{Y} = \{0, 1\}$
- Want prediction function $x \mapsto \theta = p(y = 1 \mid x)$.
- We need $\theta \in [0, 1]$.
- For a “linear method”, we can write this in two steps:

$$\underbrace{x}_{\in \mathbf{R}^d} \mapsto \underbrace{w^T x}_{\in \mathbf{R}} \mapsto \underbrace{f(w^T x)}_{\in [0,1]},$$

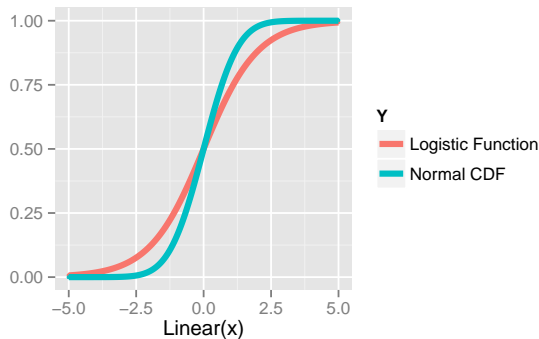
where $f : \mathbf{R} \rightarrow [0, 1]$ is called the **transfer** or **inverse link** function.

- Probability model is then

$$p(y = 1 \mid x) = f(w^T x)$$

Inverse Link Functions

- Two commonly used “inverse link” functions to map from $w^T x$ to θ :



- Logistic function \implies Logistic Regression
- Normal CDF \implies Probit Regression

Learning

- $\mathcal{X} = \mathbf{R}^d$
- $\mathcal{Y} = \{0, 1\}$
- $\mathcal{A} = , 1$ (Representing Bernoulli(θ) distributions by $\theta \in [0, 1]$)
- $\mathcal{H} = \{x \mapsto f(w^T x) \mid w \in \mathbf{R}^d\}$
- We can choose w using maximum likelihood...

Bernoulli Regression: Likelihood Scoring

- Suppose we have data $\mathcal{D} = \{(x_1, y_1), \dots, (x_n, y_n)\}$, iid.
- Compute the model likelihood for \mathcal{D} :

$$\begin{aligned} p_w(\mathcal{D}) &= \prod_{i=1}^n p_w(y_i | x_i) \text{ [by independence]} \\ &= \prod_{i=1}^n [f(w^T x_i)]^{y_i} [1 - f(w^T x_i)]^{1-y_i}. \end{aligned}$$

- Huh? Remember $y_i \in \{0, 1\}$.
- Easier to work with the log-likelihood:

$$\log p_w(\mathcal{D}) = \sum_{i=1}^n y_i \log f(w^T x_i) + (1 - y_i) \log [1 - f(w^T x_i)]$$

Bernoulli Regression: MLE

- Maximum Likelihood Estimation (MLE) finds w maximizing $\log p_w(\mathcal{D})$.
- Equivalently, minimize the objective function

$$J(w) = - \left[\sum_{i=1}^n y_i \log f(w^T x_i) + (1 - y_i) \log [1 - f(w^T x_i)] \right]$$

- For differentiable f ,
 - $J(w)$ is differentiable, and we can use our standard tools.
- Homework: Derive the SGD step directions for logistic regression.

Multinomial Logistic Regression

Multinomial Logistic Regression

- Setting: $\mathcal{X} = \mathbf{R}^d$, $\mathcal{Y} = \{1, \dots, k\}$
- The numbers $(\theta_1, \dots, \theta_k)$ where $\sum_{c=1}^k \theta_c = 1$ represent a
 - “**multinoulli**” or “**categorical**” distribution.
- For each x , we want to produce a distribution on the k classes.
- That is, for each x and each $y \in \{1, \dots, y\}$, we want to produce a probability

$$p(y | x) = \theta_y,$$

where $\sum_{y=1}^K \theta_y = 1$.

Multinomial Logistic Regression: Classic Setup

- From each x , we compute a linear score function for each class:

$$x \mapsto (\langle w_1, x \rangle, \dots, \langle w_k, x \rangle) \in \mathbf{R}^k$$

- We need to map this \mathbf{R}^k vector into a probability vector.
- Use the **softmax function**:

$$(\langle w_1, x \rangle, \dots, \langle w_k, x \rangle) \mapsto \left(\frac{\exp(w_1^T x)}{\sum_{c=1}^K \exp(w_c^T x)}, \dots, \frac{\exp(w_k^T x)}{\sum_{c=1}^K \exp(w_c^T x)} \right)$$

- If $\theta \in \mathbf{R}^k$ is the output of the softmax, note that

$$\begin{aligned} \theta_i &> 0 \\ \sum_{i=1}^k \theta_i &= 1 \end{aligned}$$

Multinomial Logistic Regression: Classic Setup

- Putting this together, we write multinomial logistic regression as

$$p(y | x) = \frac{\exp(w_y^T x)}{\sum_{c=1}^K \exp(w_c^T x)},$$

where we've introduced parameter vectors $w_1, \dots, w_k \in \mathbf{R}^d$.

- Can view $x \mapsto w_y^T x$ as the score for class y , for $y \in \{1, \dots, k\}$.
- We can also “flatten” this as we did for multiclass classification.
 - Introduce a class-sensitive feature vector $\Psi(x, y) \in \mathbf{R}^{dk}$
 - Parameter vector $w \in \mathbf{R}^{dk}$.
- The log of this likelihood is concave and straightforward to optimize.

Poisson Regression

Poisson Regression: Setup

- Input space $\mathcal{X} = \mathbf{R}^d$, Output space $\mathcal{Y} = \{0, 1, 2, 3, 4, \dots\}$
- Hypothesis space consists of functions $f : x \mapsto \text{Poisson}(\lambda(x))$.
 - That is, for each x , $f(x)$ returns a Poisson with mean $\lambda(x) \in (0, \infty)$.
 - What function?
- Recall $\lambda > 0$.
- In Poisson regression, x enters **linearly**: $x \mapsto w^T x \mapsto \lambda = f(w^T x)$.
- Standard approach is to take

$$\lambda(x) = \exp(w^T x),$$

for some parameter vector w .

- Note that range of $\lambda(x) = (0, \infty)$, (appropriate for the Poisson parameter).

Poisson Regression: Likelihood Scoring

- Suppose we have data $\mathcal{D} = \{(x_1, y_1), \dots, (x_n, y_n)\}$.
- Recall the log-likelihood for Poisson is:

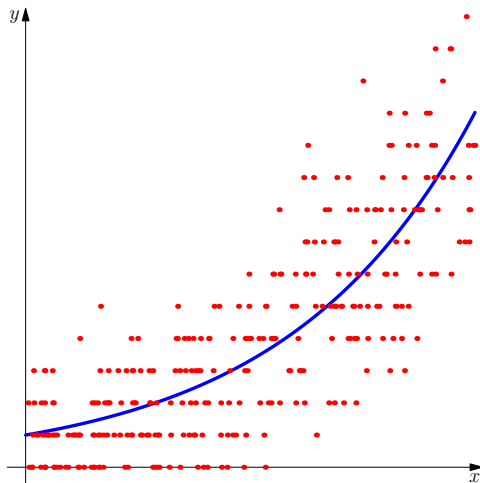
$$\log p(\mathcal{D}, \lambda) = \sum_{i=1}^n [y_i \log \lambda - \lambda - \log(y_i!)]$$

- Plugging in $\lambda(x) = \exp(w^T x)$, we get

$$\begin{aligned} \log p(\mathcal{D}, \lambda) &= \sum_{i=1}^n [y_i \log [\exp(w^T x)] - \exp(w^T x) - \log(y_i!)] \\ &= \sum_{i=1}^n [y_i w^T x - \exp(w^T x) - \log(y_i!)] \end{aligned}$$

- Maximize this w.r.t. w to find the Poisson regression.
- No closed form for optimum, but it's concave, so easy to optimize.

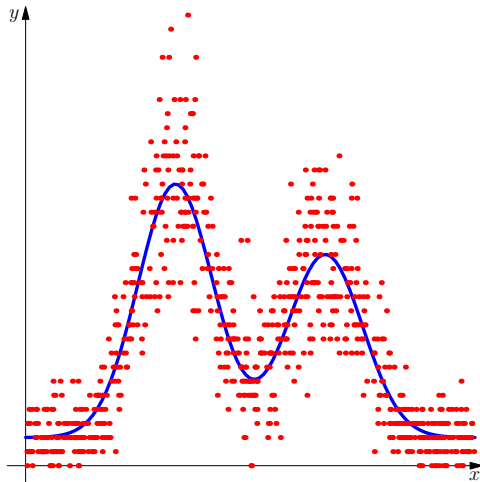
Poisson Regression Example



e.g. Phone call counts per day for a startup company over 300 days.

What About Nonlinear Score Functions

Poisson Count Example



Let's Use Gradient Boosting

- Recall the log-likelihood for Poisson regression

$$\log p(\mathcal{D}, \lambda) = \sum_{i=1}^n [y_i w^T x - \exp(w^T x) - \log(y_i!)]$$

- Let's replace $w^T x$ by a general function $f(x)$:

$$J(f) = \sum_{i=1}^n [y_i f(x_i) - \exp(f(x_i)) - \log(y_i!)]$$

Generalized Regression

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** $\mathcal{A} = \{p(y) \mid p \text{ is a probability density or mass function on } \mathcal{Y}\}$.
- Hypothesis spaces contain decision functions $f : \mathcal{X} \rightarrow \mathcal{A}$.
 - Given an $x \in \mathcal{X}$, predict a probability distribution $p(y)$ on \mathcal{Y} .

A Note on Notation

- Hypothesis spaces contain decision functions $f : \mathcal{X} \rightarrow \mathcal{A}$.
 - Given an $x \in \mathcal{X}$, predict a probability distribution $p(y)$ on \mathcal{Y} .
- Let f be a decision function.
 - In regression, $f(x) \in \mathbf{R}$
 - In hard classification, $f(x) \in \{-1, 1\}$
 - For generalized regression, $f(x) \in ?$
- $f(x)$ is a PDF or PMF on \mathcal{Y} .
- If $p = f(x)$, can evaluate $p(y)$ for predicted probability of y .
- Or just write $[f(x)](y)$ or even $f(x)(y)$.

Generalized Regression as Statistical Learning

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

$$R(f) = -\mathbb{E}_{x,y} \log [f(x)](y),$$

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

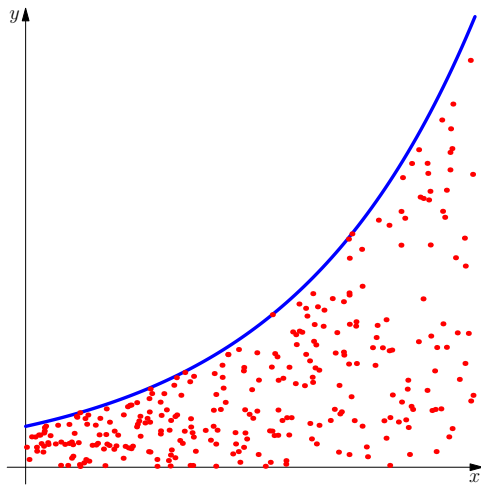
- The empirical risk of f for a sample $\mathcal{D} = \{y_1, \dots, y_n\} \in \mathcal{Y}$ is

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

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

How General A Distribution Can We Use?

Uniform Example?



- Can't use it in GBM: likelihood not differentiable (not continuous)