MLE and Conditional Probability Models

Shubham Chandel, Xintian Han & David S. Rosenberg

New York University

March 25, 2020

Contents

Maximum Likelihood

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

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

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 $\left\{\mathcal{N}(\mu, \sigma^2) \mid \mu \in \mathbf{R}, \sigma^2 > 0\right\}$
- 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.

Bernoulli Regression

Probabilistic Binary Classifiers

- Setting: $\mathcal{X} = \mathbf{R}^d$, $\mathcal{Y} = \{0, 1\}$
- For each x, need to predict a distribution on $\mathcal{Y} = \{0, 1\}$.
- How can we define a distribution supported on $\{0, 1\}$?
- Sufficient to specify the **Bernoulli parameter** $\theta = p(y = 1)$.
- We can refer to this distribution as Bernoulli(θ).

Linear Probabilistic Classifiers

- Setting: $\mathcal{X} = \mathbf{R}^d$, $\mathcal{Y} = \{0, 1\}$
- Want prediction function to map each $x \in \mathbf{R}^d$ to $\theta \in [0, 1]$.
- We first **extract information** from $x \in \mathbf{R}^d$ and summarize in a single number.
  - That number is analogous to the **score** in classification.
- For a **linear method**, this extraction is done with a linear function:

$$\underbrace{x}_{\in \mathbf{R}^d} \mapsto \underbrace{w^T x}_{\in \mathbf{R}}$$

- As usual, $x \mapsto w^T x$ will include affine functions if we include a constant feature in x.
- $w^T x$ is called the **linear predictor.**
- Still need to map this to $[0, 1]$.

The Transfer Function

- Need a function to map the linear predictor in $\mathbf{R}$ to $[0, 1]$:

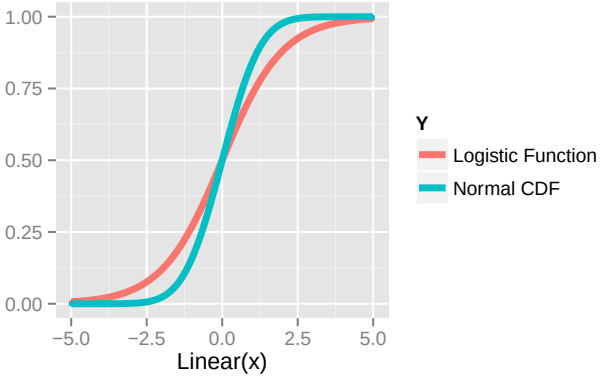
$$\underbrace{x}_{\in \mathbf{R}^d} \mapsto \underbrace{w^T x}_{\in \mathbf{R}} \mapsto \underbrace{f(w^T x)}_{\in [0,1]} = \theta,$$

  where $f : \mathbf{R} \to [0, 1]$. We'll call f the **transfer** function.

- So prediction function is $x \mapsto f(w^T x)$.

Transfer Functions for Bernoulli

- Two commonly used transfer functions to map from $w^T x$ to θ:



- Logistic function: $f(\eta) = \frac{1}{1+e^{-\eta}} \implies$ Logistic Regression
- Normal CDF $f(\eta) = \int_{-\infty}^{\eta} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \implies$ Probit Regression

Learning

- Input space $\mathcal{X} = \mathbf{R}^d$
- Outcome space $\mathcal{Y} = \{0, 1\}$
- Action space $\mathcal{A} = [0, 1]$ (Representing Bernoulli(θ) distributions by $\theta \in [0, 1]$)
- Hypothesis space $\mathcal{F} = \left\{ x \mapsto f(w^T x) \mid w \in \mathbf{R}^d \right\}$
- Parameter space $\mathbf{R}^d$ (Each prediction function represented by $w \in \mathbf{R}^d$.)
- We can choose w using maximum likelihood...

A Clever Way To Write $\hat{p}(y \mid x; w)$

- For a given $x, w \in \mathbf{R}^d$ and $y \in \{0, 1\}$, the likelihood of w for (x, y) is

$$p(y \mid x; w) = \begin{cases} f(w^T x) & y = 1 \\ 1 - f(w^T x) & y = 0 \end{cases}$$

- It will be convenient to write this as

$$p(y \mid x; w) = \left[f(w^T x)\right]^y \left[1 - f(w^T x)\right]^{1-y},$$

which is obvious as long as you remember $y \in \{0, 1\}$.

Bernoulli Regression: Likelihood Scoring

- Suppose we have data $\mathcal{D}: (x_1, y_1), \ldots, (x_n, y_n) \in \mathbf{R}^d \times \{0, 1\}$.
- The likelihood of $w \in \mathbf{R}^d$ for data $\mathcal{D}$ is

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

- Easier to work with the log-likelihood:

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

Bernoulli Regression: MLE

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

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

- For differentiable f,
  - $J(w)$ is differentiable, and we can use SGD.
  - What guarantees us to find the global minima of $J(w)$ by SGD?
  - Convexity of $J(w)$!

Multinomial Logistic Regression

Multinomial Logistic Regression

- Setting: $\mathcal{X} = \mathbf{R}^d$, $\mathcal{Y} = \{1, \ldots, k\}$
- For each x, we want to produce a distribution on k classes.
- Such a distribution is called a "**multinoulli**" or "**categorical**" distribution.
- Represent categorical distribution by probability vector $\theta = (\theta_1, \ldots, \theta_k) \in \mathbf{R}^k$:
  - $\sum_{i=1}^k \theta_i = 1$ and $\theta_i \geqslant 0$ for $i = 1, \ldots, k$ (i.e. θ represents a **distribution**) and
- So $\forall y \in \{1, \ldots, k\}$, $p(y) = \theta_y$.

Multinomial Logistic Regression

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

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

  where we've introduced parameter vectors $w_1, \ldots, w_k \in \mathbf{R}^d$.
- We need to map this $\mathbf{R}^k$ vector of scores into a probability vector.
- Consider the **softmax function:**

$$(s_1, \ldots, s_k) \mapsto \theta = \left(\frac{e^{s_1}}{\sum_{i=1}^k e^{s_i}}, \ldots, \frac{e^{s_k}}{\sum_{i=1}^k e^{s_i}} \right).$$

- Note that $\theta \in \mathbf{R}^k$ and

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

Multinomial Logistic Regression

- Say we want to get the predicted categorical distribution for a given $x \in \mathbf{R}^d$.
- First compute the scores ($\in \mathbf{R}^k$) and then their softmax:

$$x \mapsto (\langle w_1, x \rangle, \ldots, \langle w_k, x \rangle) \mapsto \theta = \left(\frac{\exp\left(w_1^T x\right)}{\sum_{i=1}^k \exp\left(w_i^T x\right)}, \ldots, \frac{\exp\left(w_k^T x\right)}{\sum_{i=1}^k \exp\left(w_i^T x\right)} \right)$$

- We can write the conditional probability for any $y \in \{1, \ldots, k\}$ as

$$p(y \mid x; w) = \frac{\exp\left(w_y^T x\right)}{\sum_{i=1}^k \exp\left(w_i^T x\right)}.$$

Multinomial Logistic Regression

- Putting this together, we write multinomial logistic regression as

$$p(y \mid x; w) = \frac{\exp\left(w_y^T x\right)}{\sum_{i=1}^{k} \exp\left(w_i^T x\right)}.$$

- How do we do learning here? What parameters are we estimating?
- Our model is specified once we have $w_1, \ldots, w_k \in \mathbf{R}^d$.
- Find parameter settings maximizing the log-likelihood of data $\mathcal{D}$.
- This objective function is concave in w's and straightforward to optimize.

Conditional Gaussian Regression

Gaussian Linear Regression

- Input space $\mathcal{X} = \mathbf{R}^d$, Output space $\mathcal{Y} = \mathbf{R}$
- In Gaussian regression, prediction functions produce a distribution $\mathcal{N}(\mu, \sigma^2)$.
  - Assume σ^2 is known.
- Represent $\mathcal{N}(\mu, \sigma^2)$ by the mean parameter $\mu \in \mathbf{R}$.
- Action space $\mathcal{A} = \mathbf{R}$
- In Gaussian linear regression, x enters **linearly:** $x \mapsto \underbrace{w^T x}_{\mathbf{R}} \mapsto \mu = \underbrace{f(w^T x)}_{\mathbf{R}}$.
- Since $\mu \in \mathbf{R}$, we can take the identity transfer function: $f(w^T x) = w^T x$.

Gaussian Regression: Likelihood Scoring

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

$$p(\mathcal{D}; w) = \prod_{i=1}^{n} p(y_i \mid x_i; w) \text{ [by independence]}$$

- Maximum Likelihood Estimation (MLE) finds w maximizing $\hat{p}(\mathcal{D}; w)$.
- Equivalently, maximize the data log-likelihood:

$$w^* = \arg\max_{w \in \mathbf{R}^d} \sum_{i=1}^{n} \log p(y_i \mid x_i; w)$$

- Let's start solving this!

Gaussian Regression: MLE

- The conditional log-likelihood is:

$$\begin{aligned}
& \sum_{i=1}^{n} \log p(y_i \mid x_i; w) \\
= & \sum_{i=1}^{n} \log \left[\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - w^T x_i)^2}{2\sigma^2} \right) \right] \\
= & \underbrace{\sum_{i=1}^{n} \log \left[\frac{1}{\sigma\sqrt{2\pi}} \right]}_{\text{independent of } w} + \sum_{i=1}^{n} \left(-\frac{(y_i - w^T x_i)^2}{2\sigma^2} \right)
\end{aligned}$$

  - MLE is the w where this is maximized.
  - Note that σ^2 is irrelevant to finding the maximizing w.
  - Can drop the negative sign and make it a minimization problem.

Gaussian Regression: MLE

- The MLE is

$$w^* = \arg\min_{w \in \mathbf{R}^d} \sum_{i=1}^{n} (y_i - w^T x_i)^2$$

- This is exactly the objective function for least squares.
- From here, can use usual approaches to solve for w^* (SGD, linear algebra, calculus, etc.)