

Review: MLE and Conditional Probability Models

Haau-Sing Li

CDS, NYU

March 10, 2019

Contents

- 1 Maximum Likelihood
- 2 Conditional Probability Models
- 3 Review Questions

Maximum Likelihood

Maximum Likelihood Estimation

- Suppose $\mathcal{D} = (y_1, \dots, y_n)$ 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

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

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).

Estimating Distributions, Overfitting, and Hypothesis Spaces

- Just as in classification and regression, MLE can overfit!

- Example Probability Models:

- $\mathcal{F} = \{\text{Poisson distributions}\}.$
- $\mathcal{F} = \{\text{Negative binomial distributions}\}.$
- $\mathcal{F} = \{\text{Histogram with 10 bins}\}$
- $\mathcal{F} = \{\text{Histogram with bin for every } y \in \mathcal{Y}\}$ [will likely overfit for continuous data]

- How to judge which model works the best?

- Choose the model with the **highest likelihood on validation set**.

$$p(y, \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Handwritten notes and diagrams: A bracket above the formula indicates n . To the right, a diagram shows a circle containing $y + \frac{1}{2}$ and $\frac{1}{2}$, with a minus sign. Below this, the formula $p^s (1-p)^{1-s}$ is written.



Conditional Probability Models

Bernoulli Regression

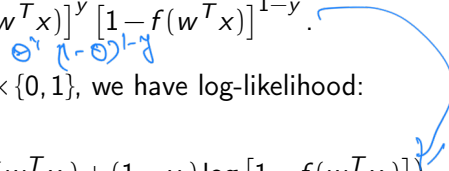
- Setting: $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \{0, 1\}$
- For each x , we predict a distribution on $\mathcal{Y} = \{0, 1\}$.
- We specify the **Bernoulli parameter** $\theta = p(y = 1)$.
- We use transfer function to map a predictor (e.g. **Linear Predictor**) to $\{0, 1\}$, referring to the Bernoulli distribution $\text{Bernoulli}(\theta)$.
- Linear Probabilistic Classifier:

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

- $w^T x$: the linear predictor; f : the **transfer** function.

Bernoulli Regression: MLE

- It will be convenient to write likelihood of w for (x, y) as this as

$$p(y | x; w) = [f(w^T x)]^y [1 - f(w^T x)]^{1-y}.$$


- With data $\mathcal{D} : (x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^d \times \{0, 1\}$, we have log-likelihood:

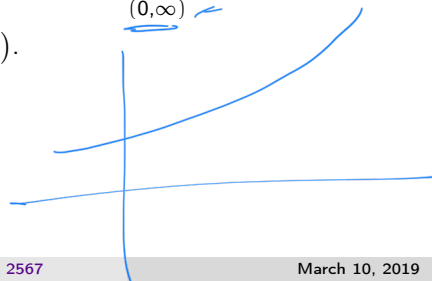
$$\log p(\mathcal{D}; w) = \sum_{i=1}^n (y_i \log f(w^T x_i) + (1 - y_i) \log [1 - f(w^T x_i)])$$

which is the negative of the **negative log-likelihood** objective $J(w)$.

- Optimization: Week 2. (Note: $J(w)$ is convex.)

Poisson Regression

- Input space $\mathcal{X} = \mathbb{R}^d$, Output space $\mathcal{Y} = \{0, 1, 2, 3, 4, \dots\}$, Action space $\mathcal{A} = (0, \infty)$.
- In Poisson regression, prediction functions produce a Poisson distribution with mean parameter $\lambda \in (0, \infty)$.
- In Poisson regression, x enters **linearly**: $x \mapsto \underbrace{w^T x}_{\mathbb{R}} \mapsto \underline{\lambda} = \underbrace{f(w^T x)}_{(0, \infty)}$.
 - standard transfer function: $f(w^T x) = \exp(w^T x)$.



Poisson Regression: MLE

\mathcal{D}

- The likelihood for w on the full dataset \mathcal{D} is

$$\log p(\mathcal{D}; w) = \sum_{i=1}^n [y_i \underbrace{w^T x_i}_{\text{mean}} - \underbrace{\exp(w^T x_i)}_{\text{exp}} - \log(y_i!)]$$

- To get MLE, need to maximize

$$J(w) = \log p(\mathcal{D}; w)$$

over $w \in \mathbb{R}^d$.

- No closed form for optimum, but it's concave, so easy to optimize.

Gaussian Linear Regression

- Input space $\mathcal{X} = \mathbb{R}^d$, Output space $\mathcal{Y} = \mathbb{R}$, Action space $\mathcal{A} = \mathbb{R}$.
- In Gaussian regression, prediction functions produce a distribution $\mathcal{N}(\mu, \sigma^2)$.
 - Assume σ^2 is known. ✓
 - We predict $\mu \in \mathbb{R}$.
- In Gaussian linear regression, x enters **linearly**: $x \mapsto \underbrace{w^T x}_{\mathbb{R}} \mapsto \mu = f(\underbrace{w^T x}_{\mathbb{R}})$.
 - Identity transfer function: $f(w^T x) = w^T x$.

Gaussian Regression: MLE

- We assume data as i.i.d. samples.
- The conditional log-likelihood is:

$$\sum_{i=1}^n \log p(y_i | x_i; w) = \text{constant} + \sum_{i=1}^n \left(-\frac{(y_i - w^T x_i)^2}{2\sigma^2} \right)$$

- The MLE is

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

- This is exactly the objective function for least squares.

Multinomial Logistic Regression

- Setting: $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \{1, \dots, k\}$
- Represent categorical distribution by probability vector $\theta = (\theta_1, \dots, \theta_k) \in \mathbb{R}^k$:
 - $\sum_{i=1}^k \theta_i = 1$ and $\theta_i \geq 0$ for $i = 1, \dots, k$ (i.e. θ represents a **distribution**)
- We follow the same steps as binomial logistic regression, except for the transfer function.
 - **Softmax Transfer Function:**

$$\frac{1}{1 + e^{-w^T x_i}}$$

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

Handwritten notes:

$$j=1 \quad \theta_k = 1 - \sum_{i=1}^{k-1} \theta_i$$

Review Questions

- **Question 1:** Suppose we have samples x_1, \dots, x_n i.i.d drawn from $\text{Bernoulli}(p)$. Find the maximum likelihood estimator of p .

$$x_i \in \{0, 1\} \forall i$$

Maximum Likelihood

$$p(x_i=1) = p$$

Solution:

- The likelihood is:

$$L(p) = \prod_{i=1}^n p^{x_i} (1-p)^{(1-x_i)}.$$

- The log-likelihood is:

$$\ell(p) = \log p \sum_{i=1}^n x_i + \log(1-p) \sum_{i=1}^n (1-x_i).$$

- Set the derivative of log-likelihood w.r.t. p to zero:

$$\frac{\partial \ell(p)}{\partial p} = \frac{\sum_{i=1}^n x_i}{p} - \frac{\sum_{i=1}^n (1-x_i)}{1-p} = 0.$$

Maximum Likelihood

- Solving the equation above, we have:

$$p = \frac{1}{n} \sum_{i=1}^n x_i.$$

- The second derivative of log-likelihood w.r.t. p is

$$\frac{\partial^2 \ell(p)}{\partial p^2} = \frac{-\sum_{i=1}^n x_i}{p^2} - \frac{\sum_{i=1}^n (1-x_i)}{(1-p)^2}.$$

- Since $p \in [0, 1]$ and $x_i \in \{0, 1\}$, the second derivative is always negative. The log-likelihood is concave. Therefore, $p = \frac{1}{n} \sum_{i=1}^n x_i$ gives us the MLE.
- A twice differentiable function of one variable is concave on an interval if and only if its second derivative is non-positive there!
- Why cannot we have the same closed form solution for logistic regression?

- **Question 2:** Suppose we have samples x_1, \dots, x_n i.i.d drawn from uniform distribution $\mathcal{U}(a, b)$. Find the maximum likelihood estimator of a and b .

Solution:

- The likelihood is:

$$L(a, b) = \prod_{i=1}^n \left(\underbrace{\frac{1}{b-a}}_{\text{Indicator for}} \mathbb{1}_{[a,b]}(x_i) \right)$$

- Let $x_{(1)}, \dots, x_{(n)}$ be the order statistics.
- The likelihood is greater than zero if and only $a < x_{(1)}$ and $b > x_{(n)}$.
- When $a < x_{(1)}$ and $b > x_{(n)}$, the likelihood is a monotonically decreasing function of $(b-a)$.
- And the smallest $(b-a)$ will be attained when $b = x_{(n)}$ and $a = x_{(1)}$.
- Therefore, $b = x_{(n)}$ and $a = x_{(1)}$ give us the MLE.

- **Question 3:** We want to fit a regression model where $Y|X = x \sim \mathcal{U}([0, e^{w^T x}])$ for some $w \in \mathbb{R}^d$. Given i.i.d. data points $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathbb{R}^d \times \mathbb{R}$, give a convex optimization problem that finds the MLE for w .

a b

Maximum Likelihood

Solution: The likelihood L is given by

$$L(w; x_1, y_1, \dots, x_n, y_n) = \prod_{i=1}^n \frac{\mathbb{1}(y_i \leq e^{w^T x_i})}{(e^{w^T x_i} - 0)}.$$

Taking logs we get

$$-\sum_{i=1}^n w^T x_i = -w^T \left(\sum_{i=1}^n x_i \right)$$

if $y_i \leq \exp(w^T x_i)$ for all i , or $-\infty$ otherwise. Thus we obtain the linear program

$$\begin{aligned} & \text{minimize} && w^T \left(\sum_{i=1}^n x_i \right) \\ & \text{subject to} && \log(y_i) \leq w^T x_i \quad \text{for } i = 1, \dots, n. \end{aligned}$$

- **Question 4:** Suppose we have input-output pairs $\{(x_1, y_1), \dots, (x_n, y_n)\}$, where $x_i \in \mathbb{R}^p$ and $y_i \in N = \{0, 1, 2, 3, \dots\}$ for $i = 1, \dots, n$. Our task is to train a Poisson regression to model the data. Assume the linear coefficients in the model is w .
 - 1 Suppose a test point x^* is orthogonal to the space generated by the training data. What is the prediction ℓ_2 regularized Poisson GLM make on the test point?
 - 2 Will the solution of the parameters \hat{w} still be sparse when we use ℓ_1 regularization?

$$\begin{aligned} f(w^T x) &= \exp(w^T x) \\ \psi(w^T x) &\in (0, \infty) \end{aligned}$$

Maximum Likelihood

- Suppose a test point x^* is orthogonal to the space generated by the training data. What is the prediction ℓ_2 regularized Poisson GLM make on the test point?

Solution: ℓ_2 penalized Poisson regression objective:

$$\hat{J}(w) = - \sum_{i=1}^n [y_i w^T x_i - \exp(w^T x_i) - \log(y_i!)] + \lambda \|w\|_2^2$$

Handwritten notes: $\hat{J}(w)$ above the first term, $+ R(w)$ above the second term, and $\lambda = f(w^T x)$ below the summation term. A large curly brace on the right groups the two terms.

From Representer Theorem, the minimizer $\hat{w} = \sum_{i=1}^n \alpha_i x_i$. The prediction is

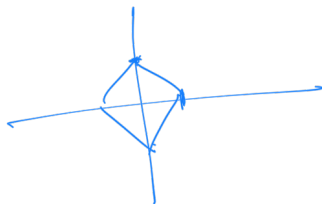
$$\exp(w^T x^*) = \exp\left(\sum_{i=1}^n \alpha_i x_i^T x^*\right) = \exp(0) = 1$$

Handwritten notes: A blue box around \hat{w} in the previous block has an arrow pointing to the w in this equation. The x_i and x^* in the summation are circled in blue. An arrow points from the summation to the 0 in $\exp(0)$, which is also circled in blue.

Maximum Likelihood

- Will the solution of the parameters \hat{w} still be sparse when we use ℓ_1 regularization?

Solution: Negative log-likelihood of Poisson regression is a convex function. The sublevel set is a convex set. The level set is the boundary of the sublevel set. When the level set approaches the diamond (level set of the ℓ_1 norm), it is still likely to hit the corner of the diamond.



- DS-GA 1003 Machine Learning Spring 2019
- DS-GA 1003 Machine Learning Spring 2020