Fundamentals of Machine Learning

MODEL FITTING & PARAMETER ESTIMATION

MAXIMUM LIKELIHOOD ESTIMATION

Amit K Roy-Chowdhury

Acknowledgments: Adapted from slides at https://probml.github.io/pml-book/teaching1.html by Prof. Saw Shier Nee
Adapted from slides of Prof. Fowler @ Binghamton University;
Modified by Prof. Roy-Chowdhury @ UC Riverside

Model Fitting / Training

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \boxed{\mathcal{L}(\boldsymbol{\theta})}$$

Loss function / objective function

Maximum likelihood estimation

$$\hat{\boldsymbol{\theta}}_{\text{mle}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{n=1}^{N} \log p(\boldsymbol{y}_n | \boldsymbol{x}_n, \boldsymbol{\theta})$$

Estimated Parameter

Probability

Model

Since most optimization algorithms are designed to minimize cost functions, we redefine the objective function to be the (conditional) negative log likelihood or NLL and we minimize NLL

$$\text{NLL}(\boldsymbol{\theta}) \triangleq -\log p(\mathcal{D}|\boldsymbol{\theta}) = -\sum_{n=1}^{N} \log p(\boldsymbol{y}_n | \boldsymbol{x}_n, \boldsymbol{\theta})$$

MLE for the Bernoulli distribution

Suppose Y is a random variable representing a coin toss, where the event Y = 1 corresponds to heads and Y = 0 corresponds to tails.
Let θ = p(Y = 1) be the probability of heads. The probability distribution for this rv is the Bernoulli.

The NLL for the Bernoulli distribution is given by

$$
\begin{aligned}
\mathrm{NLL}(\theta) &= -\log \prod_{n=1}^{N} p(y_n|\theta) \\
&= -\log \prod_{n=1}^{N} \theta^{\mathbb{I}(y_n=1)}(1-\theta)^{\mathbb{I}(y_n=0)} \\
&= -\sum_{n=1}^{N} \mathbb{I}\left(y_n = 1\right)\log\theta + \mathbb{I}\left(y_n = 0\right)\log(1-\theta) \\
&= -[N_1 \log\theta + N_0 \log(1-\theta)]
\end{aligned}
$$

MLE for the Bernoulli distribution

The MLE can be found by solving $\frac{d}{d\theta}\mathrm{NLL}(\theta) = 0$. The derivative of the NLL is

$$\frac{d}{d\theta}\mathrm{NLL}(\theta) = \frac{-N_1}{\theta} + \frac{N_0}{1-\theta} \tag{4.24}$$

and hence the MLE is given by

$$\hat{\theta}_{\mathrm{mle}} = \frac{N_1}{N_0 + N_1} \tag{4.25}$$

We see that this is just the empirical fraction of heads, which is an intuitive result.

Sufficient Statistics

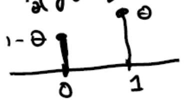
SUFFICIENT STATISTICS

A statistic T: R^k -> R^d is a sufficient statistic for $\{F_\theta\}$ if the conditional distribution of Y|T(Y) does not depend on θ.

$$P_\theta[Y^n = y^n]$$

$\to P_\theta(y_i = 1) = \theta$

$P_\theta(y_i = 0) = 1 - \theta$



$$= \prod_{i=1}^{n} \theta^{y_i} (1-\theta)^{1-y_i}$$

$$= (1-\theta)^n \left(\frac{\theta}{1-\theta}\right)^{\sum_{i=1}^{n} y_i}$$

Candidate s.s $T(y^n) = \sum_{i=1}^{n} y_i$

$$\to P_\theta\left[\widetilde{Y^n = y^n} \mid T(y^n) = k\right]$$

(y^n is such that $\sum_{i=1}^{n} y_i = k$)

$$= \frac{1}{\binom{n}{k}}$$ $\Rightarrow$ r.h.s. is ind. of θ.

MLE for Categorical Distribution

Roll a K-sided dice N times. Dataset $\mathcal{D} = \{y_n : n = 1 : N\}$

Let $Y_n \in \{1, \ldots, K\}$ be the n'th outcome, where $Y_n \sim \text{Cat}(\boldsymbol{\theta})$

$$\text{NLL}(\boldsymbol{\theta}) = -\sum_k N_k \log \theta_k \quad \text{such that} \quad \textstyle\sum_{k=1}^K \theta_k = 1$$

where N_k is the number of times the event $Y = k$ is observed.

$$\mathcal{L}(\boldsymbol{\theta}, \lambda) \triangleq -\sum_k N_k \log \theta_k - \lambda \left(1 - \sum_k \theta_k\right)$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = 1 - \sum_k \theta_k = 0$$

$$\frac{\partial \mathcal{L}}{\partial \theta_k} = -\frac{N_k}{\theta_k} + \lambda = 0 \implies N_k = \lambda \theta_k$$

$$\sum_k N_k = N_{\mathcal{D}} = \lambda \sum_k \theta_k = \lambda$$

MLE: $$\hat{\theta}_k = \frac{N_k}{\lambda} = \frac{N_k}{N}$$

Categorical Distribution - Reminder

distribution over a finite set of labels, $y \in \{1, \ldots, C\}$

$$p(y = c|\boldsymbol{\theta}) = \theta_c$$

$$\mathrm{Cat}(y|\boldsymbol{\theta}) \triangleq \prod_{c=1}^{C} \theta_c^{\mathbb{I}(y=c)}$$

$$0 \leq \theta_c \leq 1 \qquad \sum_{c=1}^{C} \theta_c = 1$$

Roll a C-sided dice N times. y is the vector that counts the number of times each face shows up.

$$y_c = N_c \triangleq \sum_{n=1}^{N} \mathbb{I}\left(y_n = c\right)$$

Distribution of y is multinomial

$$\mathcal{M}(\boldsymbol{y}|N, \boldsymbol{\theta}) \triangleq \binom{N}{y_1 \ldots y_C} \prod_{c=1}^{C} \theta_c^{y_c} = \binom{N}{N_1 \ldots N_C} \prod_{c=1}^{C} \theta_c^{N_c}$$

What happens when N=1?

Why is the categorical distribution important? Think about the output of an ML model:

$$\mathrm{Cat}(y|f(\boldsymbol{x}; \boldsymbol{\theta}))$$

Estimate Constant Signal and Variance

$x[n] = A + w[n]$ noise is $N(0,\sigma^2)$ and white

Estimate: Constant A and Noise Variance σ^2 $\Rightarrow$ $\boldsymbol{\theta} = \begin{bmatrix} A \\ \sigma^2 \end{bmatrix}$

LF is: $$p(\mathbf{x}; A, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}[x[n]-A]^2\right\}$$

Solve: $$\frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \overset{set}{=} \mathbf{0}$$

$$\frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial A} = \frac{1}{\sigma^2}\sum_{n=0}^{N-1}(x[n]-A) = \frac{N}{\sigma^2}(\bar{x}-A) = 0$$

$$\frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{n=0}^{N-1}(x[n]-A)^2 = 0$$

$$\boxed{\hat{\theta}_{ML} = \begin{bmatrix} \bar{x} \\ \frac{1}{N}\sum_n (x[n]-\bar{x})^2 \end{bmatrix}}$$

Sum of observations is the s.s. for mean; sum of squares of observations is the s.s. for the variance.

MLE for Linear Model Case

The signal model is: $\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$ with the noise $\mathbf{w} \sim N(\mathbf{0},\mathbf{C})$

Find the mean and covariance of x, write the distribution, and solve for the MLE.

Solving this gives: $\boxed{\hat{\theta}_{ML} = \left(\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{C}^{-1}\mathbf{x}}$