Parametric Methods

CAI 5107: Machine Learning

Instructor: Anowarul Kabir

Email: akabir@usf.edu

Fall 2025

Introduction

- Last class: Bayesian Decision Theory
  - Uncertainty is modeled using probability theory
  - Decision theory enables us to make optimal decision
- We will estimate these probabilities
  - Given the training set
- Approach: parametric methods

Parametric methods

- Statistical inference: make decision from the <u>samples</u> provided

- Assumption:
  - Samples are drawn from a known distribution/process (i.e. Gaussian)
- Advantage:
  - Model is defined up to a small number of parameters
  - Example: mean and variance for Gaussian
  - Once we estimate the parameters, we know the whole distribution

- <u>Maximum Likelihood Estimation (MLE)</u> to compute the parameters
  - Start with density estimation ($p(x)$)
  - Later class densities ($p(C_i|x)$) and priors ($p(C_i)$)

Maximum Likelihood Estimation (MLE)

- Assume, we have an IID sample, $\mathcal{X} = \{x^t\}_{t=1}^N$, and x_t is drawn from some known probability density, $p(x|\theta)$, where \theta are the parameters $x^t \sim p(x|\theta)$
- Target: Find \theta as likely as possible
- The likelihood of \theta given X is the product of the likelihood of the individual points (why?)

$$l(\theta|\mathcal{X}) \equiv p(\mathcal{X}|\theta) = \prod_{t=1}^{N} p(x^t|\theta)$$

- MLE: estimate \theta that makes X the most likely to be drawn

Maximum Likelihood Estimation (MLE)

- Log likelihood:

$$\mathcal{L}(\theta|\mathcal{X}) \equiv \log l(\theta|\mathcal{X}) = \sum_{t=1}^{N} \log p(x^t|\theta)$$

- Objective: maximize log likelihood
  - max $\log l(\theta|\mathcal{X})$ = min - $\log l(\theta|\mathcal{X})$
  - Or minimize –ve log likelihood
- Let's analyze some distributions to estimate parameters
  - Bernoulli distribution for 2-class problem
  - Multinomial distribution for K>2 classes
  - Gaussian (normal) distribution for class-conditional densities

Bernoulli Density

- Two possible outcomes: 0/1
- Bernoulli random variable: X
  - 1 with probability *p* if an event occurs
  - 0 with probability *1-p* if the event does not occur
  - $P(X) = p^x (1-p)^{1-x}$ where x is 0 or 1
- The expected value and variance of X are:

$$
\begin{aligned}
E[X] &= \sum_x x p(x) = 1 \cdot p + 0 \cdot (1-p) = p \\
\mathrm{Var}(X) &= \sum_x (x - E[X])^2 p(x) = p(1-p)
\end{aligned}
$$

  - Where *E(X)* is the sum of each outcome multiplied by its probability.
  - And $V(X) = E[(X - E(X))^2] = E[X^2] - E[X]^2$

Bernoulli Density (cont.)

- *p* is the only parameter to estimate.
- The log likelihood is:

$$\begin{aligned} \mathcal{L}(p|\mathcal{X}) &= \log \prod_{t=1}^{N} p^{(x^t)}(1-p)^{(1-x^t)} \\ &= \sum_t x^t \log p + \left(N - \sum_t x^t\right) \log(1-p) \end{aligned}$$

- P that maximizes the log likelihood can be found by solving dL/dp=0.

$$\hat{p} = \frac{\sum_t x^t}{N}$$

- So, MLE of the mean (E(X)) is the sample average

Multinomial Density

- Generalization of Bernoulli for *K* possible outcomes instead of 2

- Let $x_1, \ldots, x_K$ be the indicator variables where
  - x_i is 1 if the outcome is i
  - 0 otherwise

- The multinomial density is $P(x_1, x_2, \ldots, x_K) = \prod_{i=1}^{K} p_i^{x_i}$

Multinomial Density (cont.)

- For N experiments with outcomes $\mathcal{X} = \{x^t\}_{t=1}^{N}$ where

$$x_i^t = \begin{cases} 1 & \text{if experiment } t \text{ chooses state } i \\ 0 & \text{otherwise} \end{cases}$$

- The MLE of p_i is $\hat{p}_i = \frac{\sum_t x_i^t}{N}$

Gaussian (normal) Density

- Given a sample $\mathcal{X} = \{x^t\}_{t=1}^N$ with $x^t \sim \mathcal{N}(\mu, \sigma^2)$
  - Where *p(x)* = $\mathcal{N}(\mu, \sigma^2)$ is the gaussian density function
  - Mean $E[X] \equiv \mu$
  - Variance $\mathrm{Var}(X) \equiv \sigma^2$

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

- The log likelihood is

$$\mathcal{L}(\mu, \sigma | \mathcal{X}) = -\frac{N}{2}\log(2\pi) - N\log\sigma - \frac{\sum_t (x^t - \mu)^2}{2\sigma^2}$$

Gaussian Density (cont.)

- The MLE of E[X] and Var(X) are

$$
\begin{aligned}
m &= \frac{\sum_t x^t}{N} \\
s^2 &= \frac{\sum_t (x^t - m)^2}{N}
\end{aligned}
$$

Reading materials

- Chapter 4: Parametric Methods
  - 4.1 - 4.2