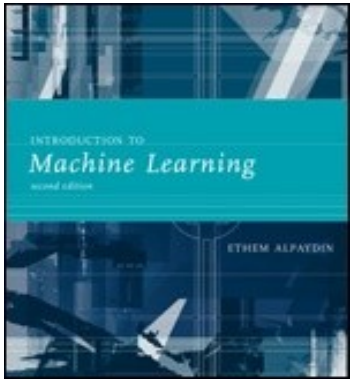
Lecture Slides for

INTRODUCTION TO

Machine Learning

2nd Edition



ETHEM ALPAYDIN, modified by Leonardo Bobadilla and some parts from http://www.cs.tau.ac.il/~apartzin/MachineLearning/

alpaydin@boun.edu.tr
http://www.cmpe.boun.edu.tr/~ethem/i2m

Outline

Last Class: Ch 3 Bayesian Decision Theory

Bayesian decision theory
Losses and risks
Discriminant functions
Association Rules

This class: CHAPTER 4: Parametric Methods

Maximum Likelihood Estimation
Evaluating an estimator: Bias and Variance

CHAPTER 4:

Parametric Methods

Parametric Methods

- Need a probabilities to make decisions (prior, evidence, likelihood)
- Probability is a function of input (observables)
- Represent function by
  - Selecting its general form (model) with several unknown parameters
  - Find(estimate) parameters from data that optimize certain criteria (e.g. minimize generalization error)

Parametric Estimation

- Assume sample comes from a distribution known up to its parameters
- Sufficient statistics : parameters that completely define distribution (e.g. mean and variance)
- $X = \{ x^t \}_t$ where $x^t \sim p(x)$
- Parametric estimation:

  Assume a form for $p(x \mid \theta)$ and estimate θ, its sufficient statistics, using X

  e.g., $N(\mu, \sigma^2)$ where $\theta = \{ \mu, \sigma^2 \}$

Estimation

- Assume form of distribution
- Estimate its sufficient parameters from a sample
- Use distribution in classification or regressions

Maximum Likelihood Estimation

- X consists of independent and identically distributed (iid) samples
- Likelihood of θ given the sample X

$$l(\theta|X) = p(X|\theta) = \prod_t p(x^t|\theta)$$

- Log likelihood

$$L(\theta|X) = \log l(\theta|X) = \sum_t \log p(x^t|\theta)$$

- Maximum likelihood estimator (MLE)

$$\theta^* = \text{argmax}_\theta L(\theta|X)$$

Why to use log likelihood

- Log is increasing function
  - Increased input->increased output
- Maximizing log of a function is equivalent to maximizing function itself
- Log convert products to sum
  - log (abc)=log(a)+log(b)+logc
- Makes analysis/computation simpler

Examples: Bernoulli/Multinomial

- Bernoulli: Two states, failure/success, *x* in {0,1}

$$\mathcal{L}(p|\mathcal{X}) = \log \prod_{t=1}^{N} p^{(x^t)}(1-p)^{(1-x^t)}$$

$$= \sum_t x^t \log p + \left(N - \sum_t x^t\right) \log(1-p)$$

  - Solving ∂L(p|X)/dp=0 => $\hat{p} = \frac{\sum_t x^t}{N}$

Examples: Multinomial

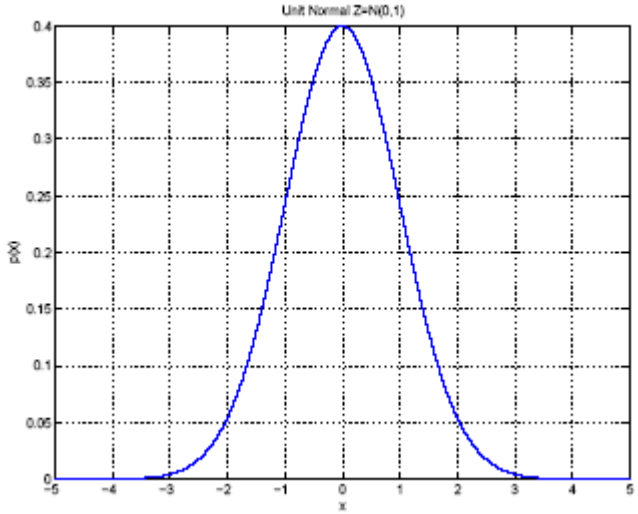
- *Generalization of Bernoulli : multinomial*
- *Outcome is mutually exclusive K states*
- *Probability of occurring* p_i $\sum_{i=1}^{K} p_i = 1$

$P(x_1,x_2,...,x_K) = \prod_i p_i^{x_i}$

$L(p_1,p_2,...,p_K|X) = \log \prod_t \prod_i p_i^{x_i^t}$

MLE: $p_i = \sum_t x_i^t / N$

Gaussian (Normal) Distribution



- $p(x) = \mathrm{N}(\mu, \sigma^2)$

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

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

$$m = \frac{\sum_t x^t}{N}$$

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

Bias and Variance of an estimator

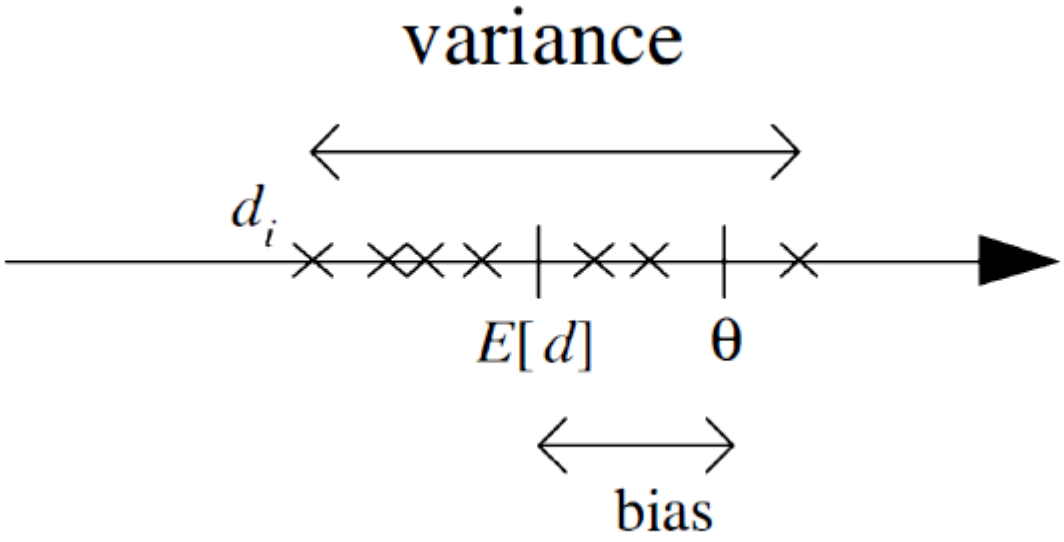
- Actual value of en estimator depends on data
- Get different N samples from a true distribution , results in different value of an estimator
- Value of an estimator is a random variables
- Can ask about its mean and variance
- Difference between the true value of a parameter and mean of estimator is bias of the estimator
- Usually looking for formula resulting in unbiased estimator with small variance

Bias and Variance

Unknown parameter θ
Estimator $d_i = d\ (X_i)$ on sample X_i

Bias: $b_\theta(d) = E\ [d] - \theta$

Bias and Variance



Mean Estimator

$$E[m] = E\left[\frac{\sum_t x^t}{N}\right] = \frac{1}{N}\sum_t E[x^t] = \frac{N\mu}{N} = \mu$$

$$\mathrm{Var}(m) = \mathrm{Var}\left(\frac{\sum_t x^t}{N}\right) = \frac{1}{N^2}\sum_t \mathrm{Var}(x^t) = \frac{N\sigma^2}{N^2} = \frac{\sigma^2}{N}$$

Variance Estimator

$$s^2 = \frac{\sum_t (x^t - m)^2}{N} = \frac{\sum_t (x^t)^2 - Nm^2}{N}$$

$$E[s^2] = \frac{\sum_t E[(x^t)^2] - N \cdot E[m^2]}{N}$$

$$E[s^2] = \frac{N(\sigma^2 + \mu^2) - N(\sigma^2/N + \mu^2)}{N} = \left(\frac{N-1}{N}\right)\sigma^2 \neq \sigma^2$$

Bias and Variance

Variance: $E\,[(d-E\,[d])^2]$

Mean square error:

$r\,(d,\theta) = E\,[(d-\theta)^2] = (E\,[d] - \theta)^2 + E\,[(d-E\,[d])^2]$

$= \text{Bias}^2 + \text{Variance}$