# Machine Learning Basics: Estimators, Bias and Variance

Sargur N. Srihari srihari@cedar.buffalo.edu

This is part of lecture slides on <a href="Deep Learning">Deep Learning</a>: http://www.cedar.buffalo.edu/~srihari/CSE676

# Topics in Basics of ML

- 1. Learning Algorithms
- 2. Capacity, Overfitting and Underfitting
- 3. Hyperparameters and Validation Sets
- 4. Estimators, Bias and Variance
- 5. Maximum Likelihood Estimation
- 6. Bayesian Statistics
- 7. Supervised Learning Algorithms
- 8. Unsupervised Learning Algorithms
- 9. Stochastic Gradient Descent
- 10. Building a Machine Learning Algorithm
- 11. Challenges Motivating Deep Learning

# Topics in Estimators, Bias, Variance

- 0. Statistical tools useful for generalization
- 1. Point estimation
- 2. Bias
- 3. Variance and Standard Error
- 4. Bias-Variance tradeoff to minimize MSE
- 5. Consistency

# Statistics provides tools for ML

- The field of statistics provides many tools to achieve the ML goal of solving a task not only on the training set but also to generalize
- Foundational concepts such as
  - Parameter estimation
  - Bias
  - Variance
- They characterize notions of generalization, over- and under-fitting

### Point Estimation

- Point Estimation is the attempt to provide the single best prediction of some quantity of interest
  - Quantity of interest can be:
    - A single parameter
    - A vector of parameters
      - E.g., weights in linear regression
    - A whole function

#### Point estimator or Statistic

- To distinguish estimates of parameters from their true value, a point estimate of a parameter  $\theta$  is represented by  $\hat{\theta}$
- Let  $\{x^{(1)}, x^{(2)}, ...x^{(m)}\}$  be m independent and identically distributed data points
  - Then a point estimator or statistic is any function of the data

$$\hat{\boldsymbol{\theta}}_m = g(\boldsymbol{x}^{(1)}, ... \boldsymbol{x}^{(m)})$$

- Thus a statistic is any function of the data
- It need not be close to the true  $\theta$
- A good estimator is a function whose output is close to the true underlying  $\theta$  that generated the data

#### **Function Estimation**

- Point estimation can also refer to estimation of relationship between input and target variables
  - Referred to as function estimation
- Here we predict a variable y given input x
  - We assume f(x) is the relationship between x and y
    - We may assume  $y=f(x)+\epsilon$ 
      - Where  $\epsilon$  stands for a part of y not predictable from x
  - We are interested in approximating f with a model  $\hat{f}$ 
    - Function estimation is same as estimating a parameter  $\theta$ 
      - where  $\hat{f}$  is a point estimator in function space
    - Ex: in polynomial regression we are either estimating a parameter  $m{w}$  or estimating a function mapping from  $m{x}$  to  $m{y}$

## Properties of Point Estimators

- Most commonly studied properties of point estimators are:
  - 1. Bias
  - 2. Variance
- They inform us about the estimators

#### 1. Bias of an estimator

• The bias of an estimator  $\hat{\theta}_m = g(\mathbf{x}^{(1)},...\mathbf{x}^{(m)})$  for parameter  $\theta$  is defined as

$$\left| \operatorname{bias}(\hat{\boldsymbol{\theta}}_m) = E[\hat{\boldsymbol{\theta}}_m] - \boldsymbol{\theta} \right|$$

- The estimator is unbiased if  $bias(\hat{\theta}_m)=0$ 
  - which implies that  $E[\hat{\theta}_m] = \theta$
- · An estimator is asymptotically unbiased if

$$\lim_{m\to\infty} \operatorname{bias}(\hat{\boldsymbol{\theta}}_m) = 0$$

# **Examples of Estimator Bias**

- We look at common estimators of the following parameters to determine whether there is bias:
  - Bernoulli distribution: mean θ
  - Gaussian distribution: mean μ
  - Gaussian distribution: variance  $\sigma^2$

#### Estimator of Bernoulli mean

- Bernoulli distribution for binary variable  $x \in \{0,1\}$  with mean  $\theta$  has the form  $P(x;\theta) = \theta^x (1-\theta)^{1-x}$
- Estimator for  $\theta$  given samples  $\{x^{(1)},...x^{(m)}\}$  is  $\hat{\theta}_m = \frac{1}{m}\sum_{i=1}^m x^{(i)}$
- To determine whether this estimator is biased

determine

$$\begin{aligned} \operatorname{bias}(\hat{\boldsymbol{\theta}}_{m}) &= E\left[\hat{\boldsymbol{\theta}}_{m}\right] - \boldsymbol{\theta} \\ &= E\left[\frac{1}{m}\sum_{i=1}^{m}x^{(i)}\right] - \boldsymbol{\theta} \\ &= \frac{1}{m}\sum_{i=1}^{m}E\left[x^{(i)}\right] - \boldsymbol{\theta} \\ &= \frac{1}{m}\sum_{i=1}^{m}\sum_{x^{(i)}=0}^{1}\left(x^{(i)}\boldsymbol{\theta}^{x^{(i)}}(1-\boldsymbol{\theta})^{(1-x^{(i)})}\right) - \boldsymbol{\theta} \\ &= \frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{\theta}) - \boldsymbol{\theta} = \boldsymbol{\theta} - \boldsymbol{\theta} = 0 \end{aligned}$$

- Since  $bias(\hat{\theta}_m)=0$  we say that the estimator is unbiased

#### Estimator of Gaussian mean

- Samples  $\{x^{(1)},...x^{(m)}\}$  are independently and identically distributed according to  $p(x^{(i)})=N(x^{(i)};\mu,\sigma^2)$ 
  - Sample mean is an estimator of the mean parameter

$$\hat{\mu}_m = \frac{1}{m} \sum_{i=1}^m x^{(i)}$$

– To determine bias of the sample mean:

$$bias(\hat{\mu}_m) = \mathbb{E}[\hat{\mu}_m] - \mu$$

$$= \mathbb{E}\left[\frac{1}{m} \sum_{i=1}^m x^{(i)}\right] - \mu$$

$$= \left(\frac{1}{m} \sum_{i=1}^m \mathbb{E}\left[x^{(i)}\right]\right) - \mu$$

$$= \left(\frac{1}{m} \sum_{i=1}^m \mu\right) - \mu$$

$$= \mu - \mu = 0$$

 Thus the sample mean is an unbiased estimator of the Gaussian mean

#### Estimator for Gaussian variance

- The sample variance is  $\left| \hat{\sigma}_m^2 = \frac{1}{m} \sum_{i=1}^m \left( x^{(i)} \hat{\mu}_m \right)^2 \right|$
- · We are interested in computing

$$\operatorname{bias}(\hat{\sigma}_{m}^{2}) = \operatorname{E}(\hat{\sigma}_{m}^{2}) - \sigma^{2}$$

- We begin by evaluating ->
- Thus the bias of  $\hat{\sigma}_m^2$  is  $-\sigma^2/m$  =  $\frac{m-1}{m}\sigma^2$

$$\mathbb{E}[\hat{\sigma}_m^2] = \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^m \left(x^{(i)} - \hat{\mu}_m\right)^2\right]$$
$$= \frac{m-1}{m}\sigma^2$$

- Thus the sample variance is a biased estimator
- The unbiased sample variance estimator is

$$\hat{\sigma}_{m}^{2} = \frac{1}{m-1} \sum_{i=1}^{m} (x^{(i)} - \hat{\mu}_{m})^{2}$$

#### 2. Variance and Standard Error

- Another property of an estimator:
  - How much we expect the estimator to vary as a function of the data sample
- Just as we computed the expectation of the estimator to determine its bias, we can compute its variance
- The variance of an estimator is simply  $Var(\hat{\theta})$  where the random variable is the training set
- The square root of the the variance is called the standard error, denoted  $SE(\hat{\theta})$

# Importance of Standard Error

- It measures how we would expect the estimate to vary as we obtain different samples from the same distribution
- The standard error of the mean is given by

$$SE(\hat{\mu}_m) = \sqrt{Var\left[\frac{1}{m}\sum_{i=1}^m x^{(i)}\right]} = \frac{\sigma}{\sqrt{m}}$$

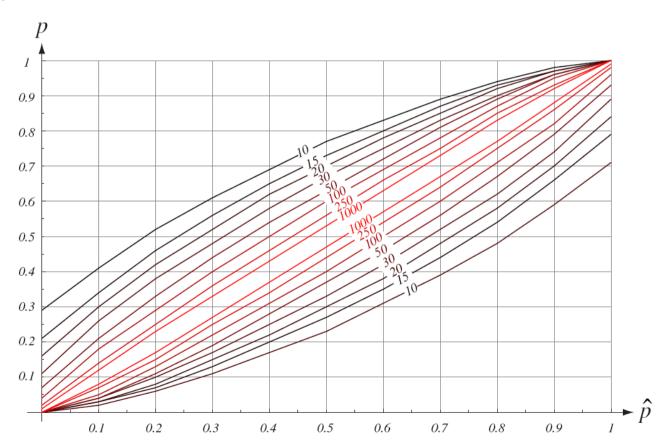
- where  $\sigma^2$  is the true variance of the samples  $\boldsymbol{x}^{(i)}$
- Standard error often estimated using estimate of  $\sigma$ 
  - Although not unbiased, approximation is reasonable
    - The standard deviation is less of an underestimate than variance

# Standard Error in Machine Learning

- We often estimate generalization error by computing error on the test set
  - No of samples in the test set determine its accuracy
  - Since mean will be normally distributed, (according to Central Limit Theorem), we can compute probability that true expectation falls in any chosen interval
    - Ex: 95% confidence interval centered on mean  $\hat{\mu}_m$  is  $\left[ \left( \hat{\mu}_m 1.96SE \left( \hat{\mu}_m \right), \hat{\mu}_m + 1.96SE \left( \hat{\mu}_m \right) \right) \right]$
- ML algorithm A is better than ML algorithm B if
  - upperbound of A is less than lower bound of B

#### Confidence Intervals for error

#### 95% confidence intervals for error estimate

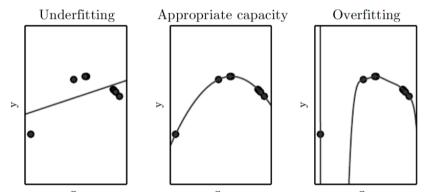


# Trading-off Bias and Variance

- Bias and Variance measure two different sources of error of an estimator
- Bias measures the expected deviation from the true value of the function or parameter
- Variance provides a measure of the expected deviation that any particular sampling of the data is likely to cause

# Negotiating between bias - tradeoff

 How to choose between two algorithms, one with a large bias and another with a large variance?



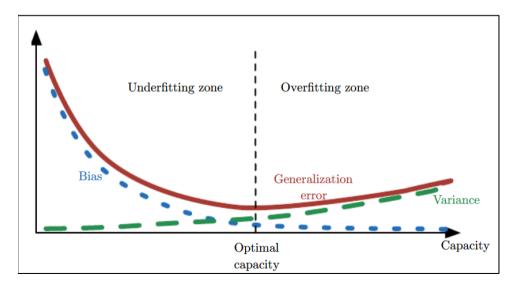
- Most common approach is to use cross-validation
- Alternatively we can minimize Mean Squared Error which incorporates both bias and variance

# Mean Squared Error

Mean Squared Error of an estimate is

$$MSE = E\left[\left(\hat{\boldsymbol{\theta}}_{m} - \boldsymbol{\theta}\right)^{2}\right]$$
$$= Bias\left(\hat{\boldsymbol{\theta}}_{m}\right)^{2} + Var\left(\hat{\boldsymbol{\theta}}_{m}\right)$$

Minimizing the MSE keeps both bias and variance in check

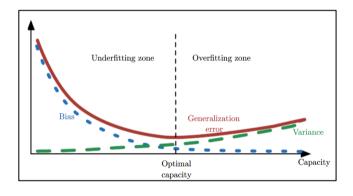


As capacity increases, bias (dotted) tends to decrease and variance (dashed) tends to increase

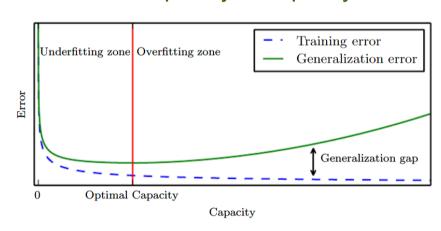
#### **Underfit-Overfit: Bias-Variance**

Relationship of bias-variance to capacity is similar to underfitting and overfitting relationship to capacity

#### Bias-Variance to capacity



#### Model complexity to capacity



Both have a U-shaped curve of generalization Error as a function of capacity

# Consistency

- So far we have discussed behavior of an estimator for a fixed training set size
- We are also interested with the behavior of the estimator as training set grows
- As the no. of data points m in the training set grows, we would like our point estimates to converge to the true value of the parameters:

$$\operatorname{plim}_{m\to\infty}\hat{\boldsymbol{\theta}}_m = \boldsymbol{\theta}$$

Symbol plim indicates convergence in probability

# Weak and Strong Consistency

•  $\operatorname{plim}_{m\to\infty}\hat{\boldsymbol{\theta}}_m = \boldsymbol{\theta}$  means that

For any 
$$\varepsilon > 0$$
,  $P(|\hat{\theta}_m - \theta| > \varepsilon) \to 0$  as  $m \to \infty$ 

- It is also known as weak consistency
  - Implies almost sure convergence of  $\hat{\theta}$  to  $\theta$
- Strong consistency refers to almost sure convergence of a sequence of random variables  $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots$  to a value  $\boldsymbol{x}$  occurs when

$$p(\lim_{m\to\infty} \boldsymbol{x}^{(m)} = \boldsymbol{x}) = 1$$

• Consistency ensures that the bias induced by the estimator decreases with m