Model Complexity
Bias and Variance Tradeoff
Error Measurement

CSE512 – Machine Learning, Spring 2018, Stony Brook University

Instructor: Minh Hoai Nguyen (minhhoai@cs.stonybrook.edu)

Date: 31 Jan 2018

Outline

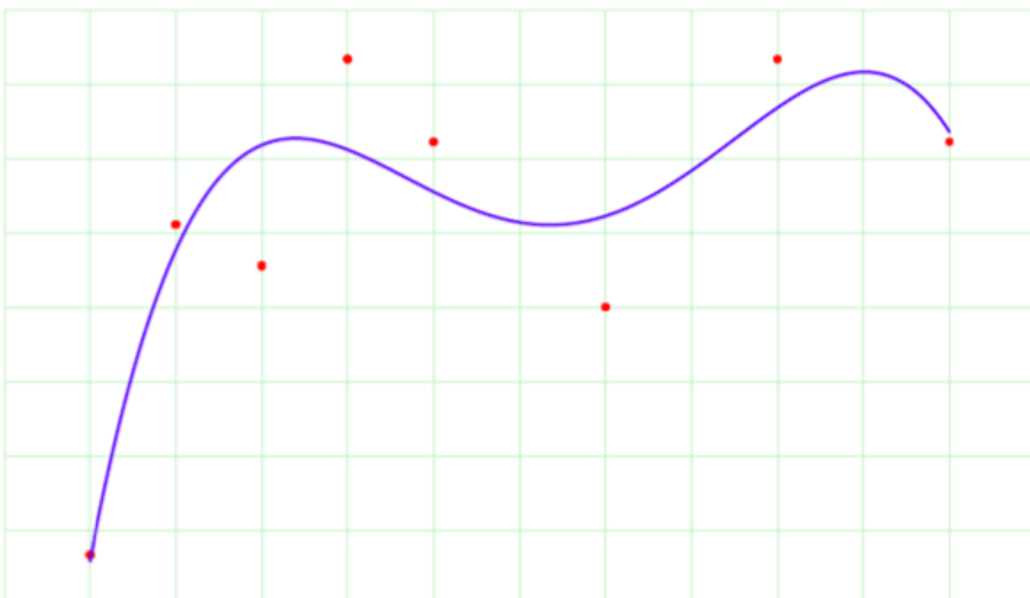
- Bias and Variance of Learner
- Train and Prediction Errors
- Common Error Measurements

Linear Regression Reviewed

Assume the output is a linear function of input features

$$\hat{y} = \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_d x_d + \theta_{d+1}$$

Suppose the output variable is a polynomial of degree 4 of the input can we use Linear Regression to learn the function?



Basic Functions

Still a linear regression problem

$$\hat{y} = \theta_1 u_1(\mathbf{x}) + \theta_2 u_2(\mathbf{x}) + \cdots + \theta_d u_d(\mathbf{x}) + \theta_{d+1}$$

Example

$$x \to (x, x^2, x^3)$$

$$(x_1, x_2) \to (x_1^2, x_2^2, x_1 x_2, x_1, x_2)$$

Model Complexity

There are different models to relate input to output:

- E.g., linear, quadratic, cubic, etc.
- There are `simple' and 'complex' models

Model too "simple":

- Does not fit the data well
- A high-bias solution

Model too "complex":

- Small changes to the data leads to large changes in the solution
- A high-variance solution

Different data lead to different solution

- Given dataset *D* with *m* samples, learn function *h*(x)
- If you sample a different dataset *D*, you will learn different *h*(x)

- Expected hypothesis: $E_D[h(x)]$

Squared Bias of Learner

- Expected hypothesis: $E_D[h(x)]$

- Bias: difference between what you expect to learn and the truth *t*(x)

$$bias^2 = \int_x (E_D[h(x)] - t(x))^2 p(x) dx$$

- Measures how well you expect to represent true solution

Squared Bias of Learner: Decrease with more complex model

$$bias^2 = \int_x (E_D[h(x)] - t(x))^2 p(x) dx$$

Variance of Learner

- Given dataset *D* with *m* samples, learn function *h(*x*)*
- If you sample a different dataset *D*, you will learn different *h*(x)

- Variance: difference between what you expect to learn and what you learn from a particular dataset

$$\bar{h}(x) = E_D[h(x)]$$

$$variance = \int_x E_D[(h(x) - \bar{h}(x))^2]p(x)dx$$

- Measures how sensitive learner is to specific dataset

Variance of Learner: Decreases with simpler model

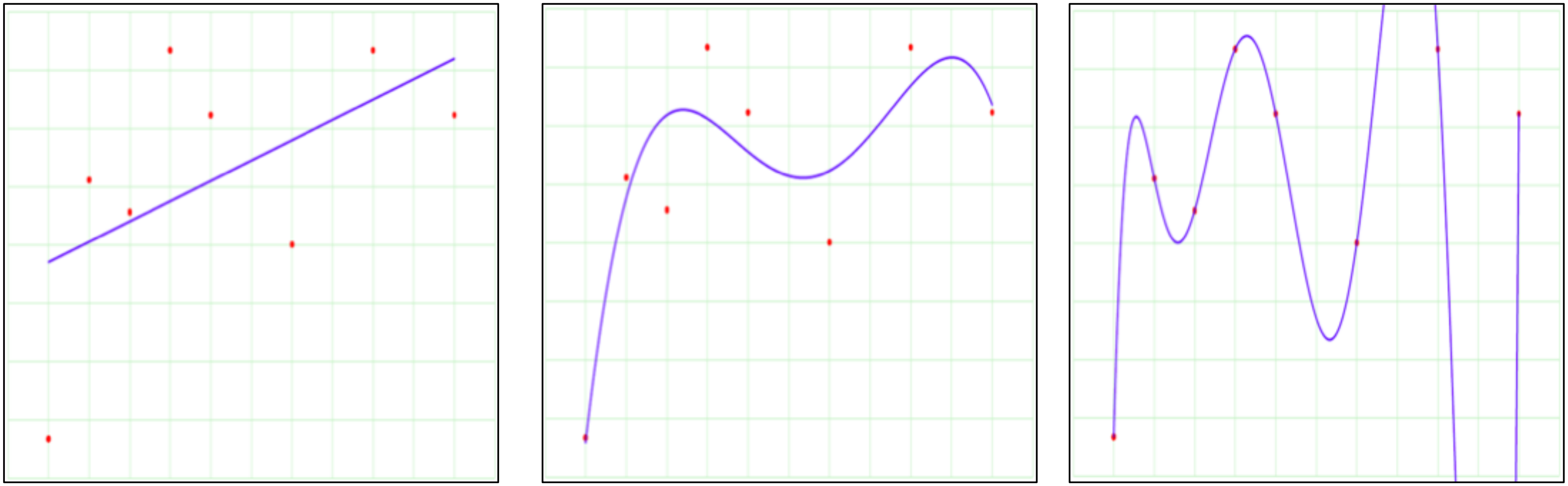
$$\bar{h}(x) = E_D[h(x)]$$

$$variance = \int_x E_D[(h(x) - \bar{h}(x))^2]p(x)dx$$

Bias-Variance Tradeoff

Choice of hypothesis class introduces learning bias and variance

- More complex class -> less bias
- More complex class -> more variance



Bias-Variance Error Decomposition

Consider the regression problem:

$$t = f(x) = g(x) + \epsilon$$

True

Deterministic

Noise $\mathcal{N}(0, \sigma)$

Task: given some training data, we learn a function h(x)

What are the sources of prediction error?

Source of Error 1 - Noise

Even when we have a perfect learner with infinite training data

- If our solution h(x) satisfies h(x) = g(x)
- We still have unavoidable error due to noise

$$error(h) = \int_x \int_t (h(x) - t)^2 p(f(x) = t|x)p(x)dtdx$$

$$= \int_x \int_\epsilon \epsilon^2 p(\epsilon)p(x)d\epsilon dx$$

$$= \sigma^2$$

Source of Error 2 – Finite Data

We have imperfect learner, or only *m* training examples

The expected squared error per example (Expectation over random training set *D* of size *m*, drawn from distribution p(x, t)

$$error(h) = E_D \left[\int_x \int_t (h(x) - t)^2 p(f(x) = t|x) p(x) dt dx \right]$$

$$= unavoidableError + variance + bias^2$$

With

$$unavoidableError = \sigma^2$$

$$bias^2 = \int_x (E_D[h(x)] - t(x))^2 p(x) dx$$

$$\bar{h}(x) = E_D[h(x)]$$

$$variance = \int_x E_D[(h(x) - \bar{h}(x))^2] p(x) dx$$

Outline

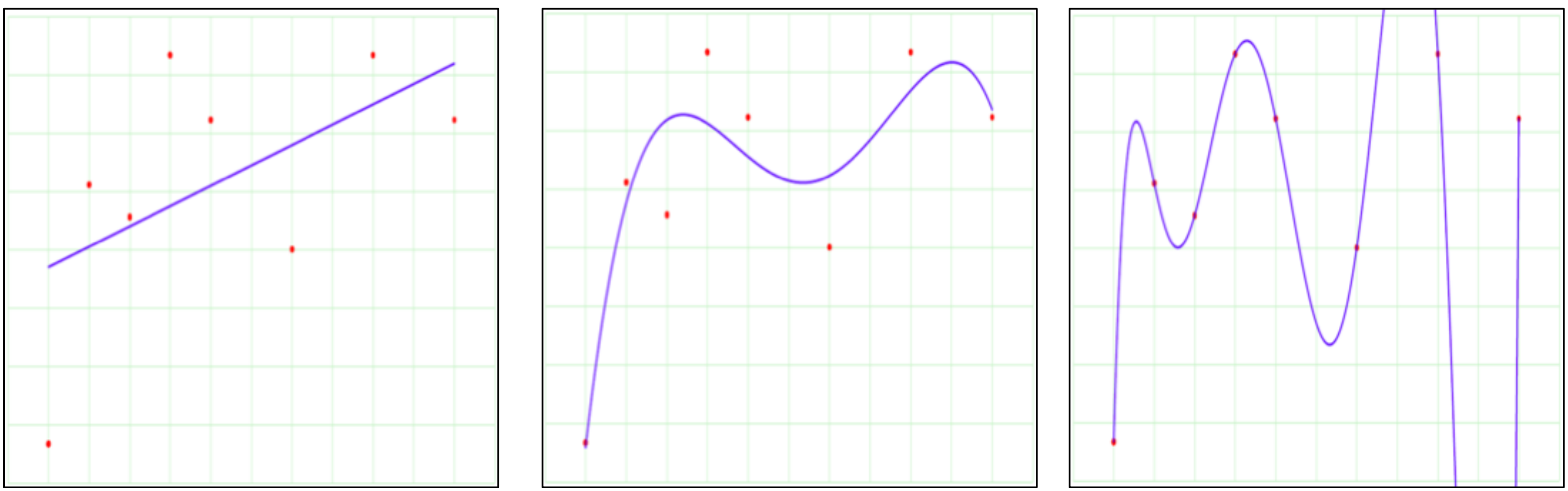
- Bias and Variance of Learner
- Train and Prediction Errors
- Common Error Measurements

Bias-Variance Tradeoff

Choice of hypothesis class introduces learning bias and variance

- More complex class -> less bias
- More complex class -> more variance

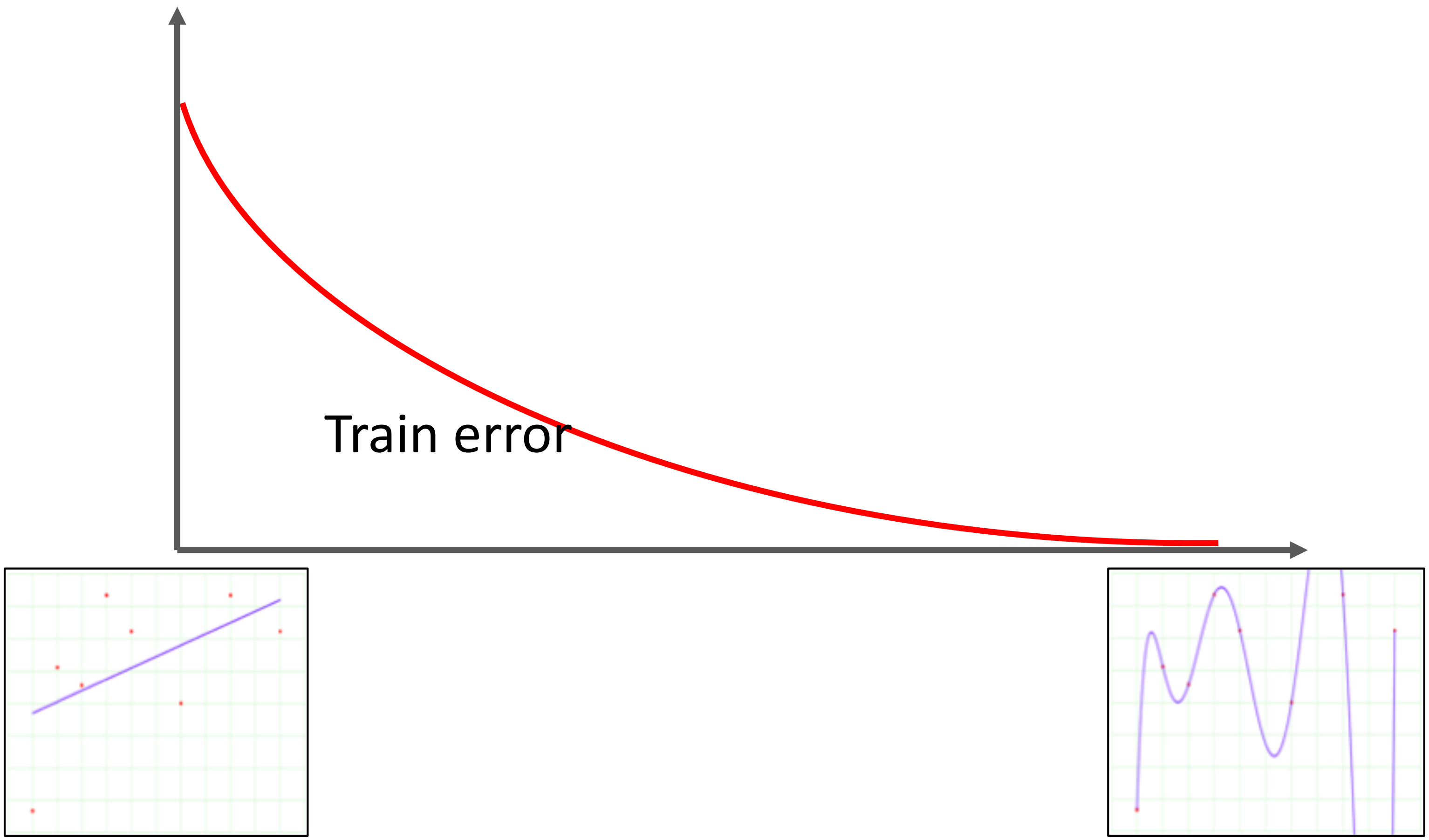
Underfitting Overfitting



Training Set Error

Measure on the training data:

$$Error_{train}(\boldsymbol{\theta}) = \frac{1}{n_{train}} \sum_{i=1}^{n_{train}} (y_i - \boldsymbol{\theta}^T \mathbf{x}_i)^2$$



Prediction Error

We care about error over all possible input points, not just training data

$$Error_{true}(\boldsymbol{\theta}) = \int_x (y^*(\mathbf{x}) - \boldsymbol{\theta}^T\mathbf{x})^2 P(x)dx$$

Prediction Error

We care about error over all possible input points, not just training data

$$Error_{true}(\boldsymbol{\theta}) = \int_x (y^*(\mathbf{x}) - \boldsymbol{\theta}^T \mathbf{x})^2 P(x) dx$$

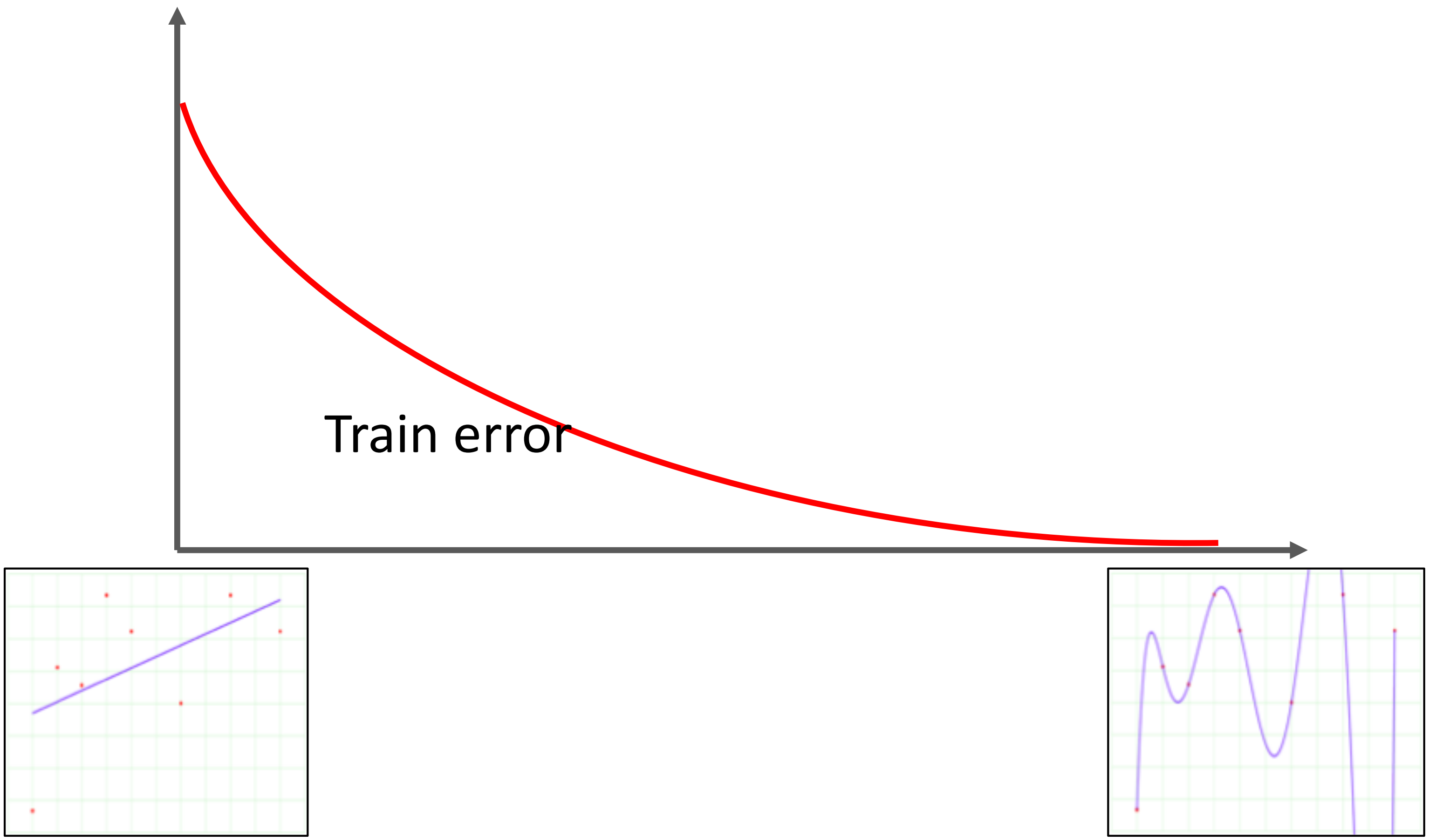
Training error is NOT a good estimate for true error:

- Because we cheated
- We used training data to select parameters with lowest error
- Training error is an optimistically biased estimate of prediction error

Test Set Error

Measure error on an independent test set instead!

Train and test error as a function of model complexity



Error as a function of number of training examples for a fixed model complexity

Little data

Infinite data

Warning

- Test set only unbiased if you NEVER NEVER NEVER NEVER NEVER NEVER NEVER NEVER NEVER do ANY learning on the test data
- E.g., you cannot use test set to select the degree of the polynomial or the regularization parameter

Outline

- Bias and Variance of Learner
- Train and Prediction Errors
- Common Error Measurements

Root Mean Squared Error (RMSE) for Regression

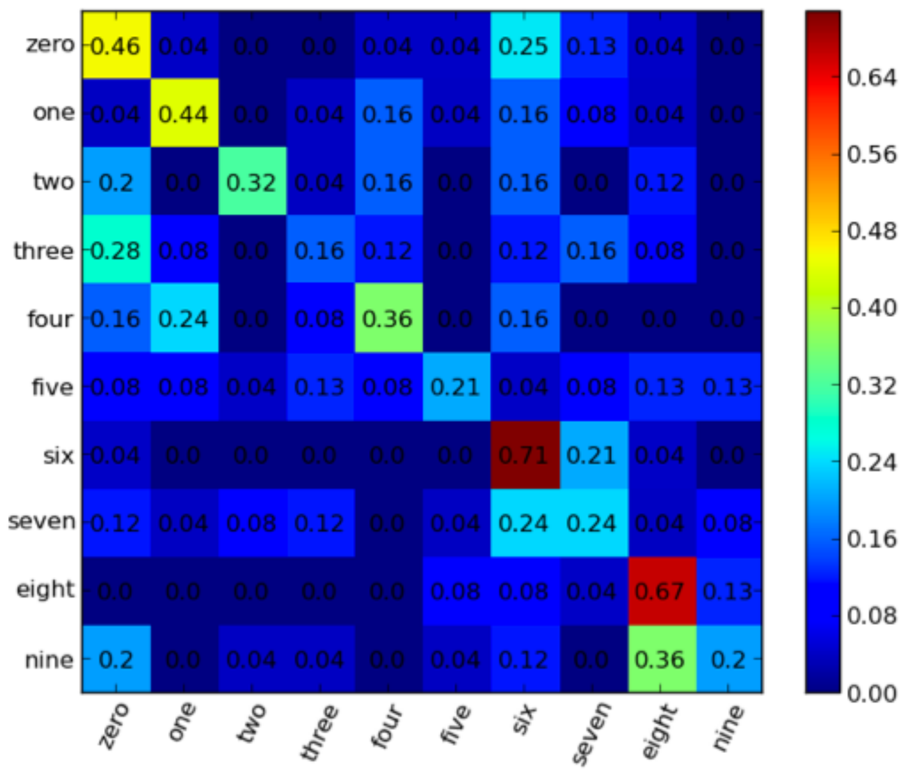
$$Error_{test}(h) = \sqrt{\frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (y_i - h(\mathbf{x}_i))^2}$$

Accuracy for Classification Problem

$$Accuracy_{test}(h) = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \delta(y_i = h(\mathbf{x}_i))$$

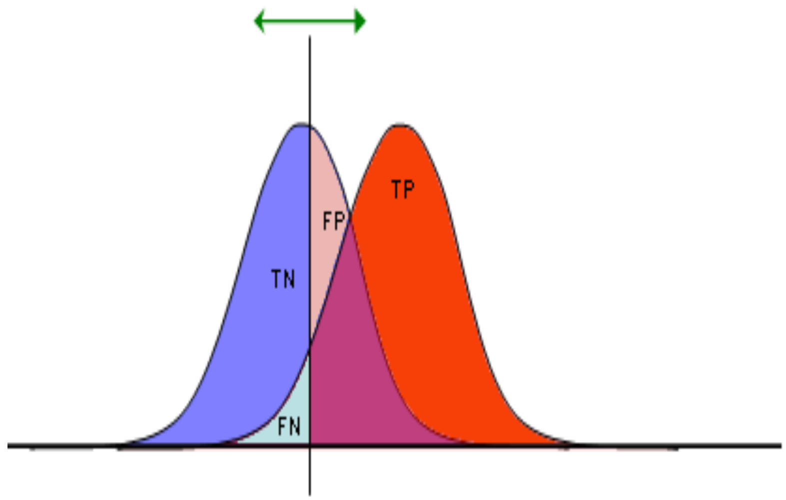
$$Error_{test}(h) = 1 - Accuracy_{test}(h)$$

Confusion Matrix



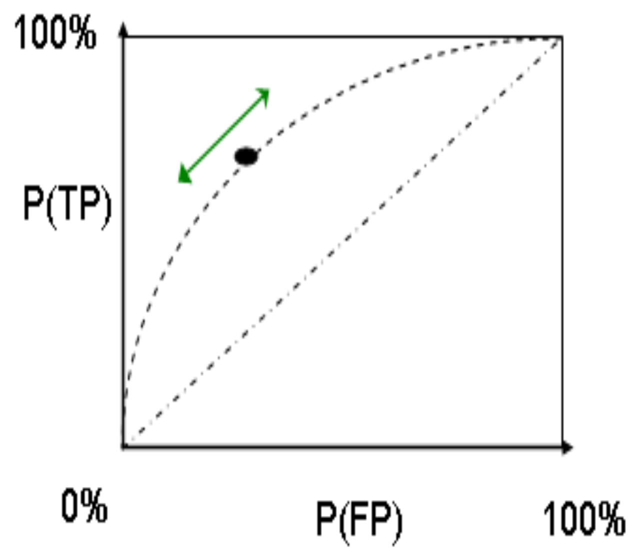
Receiver Operating Characteristic (ROC)

Consider a binary classifier: X is classified as positive iff h(X) > θ



TP	FP
FN	TN
1	1

Sensitivity,
Recall
True positive rate

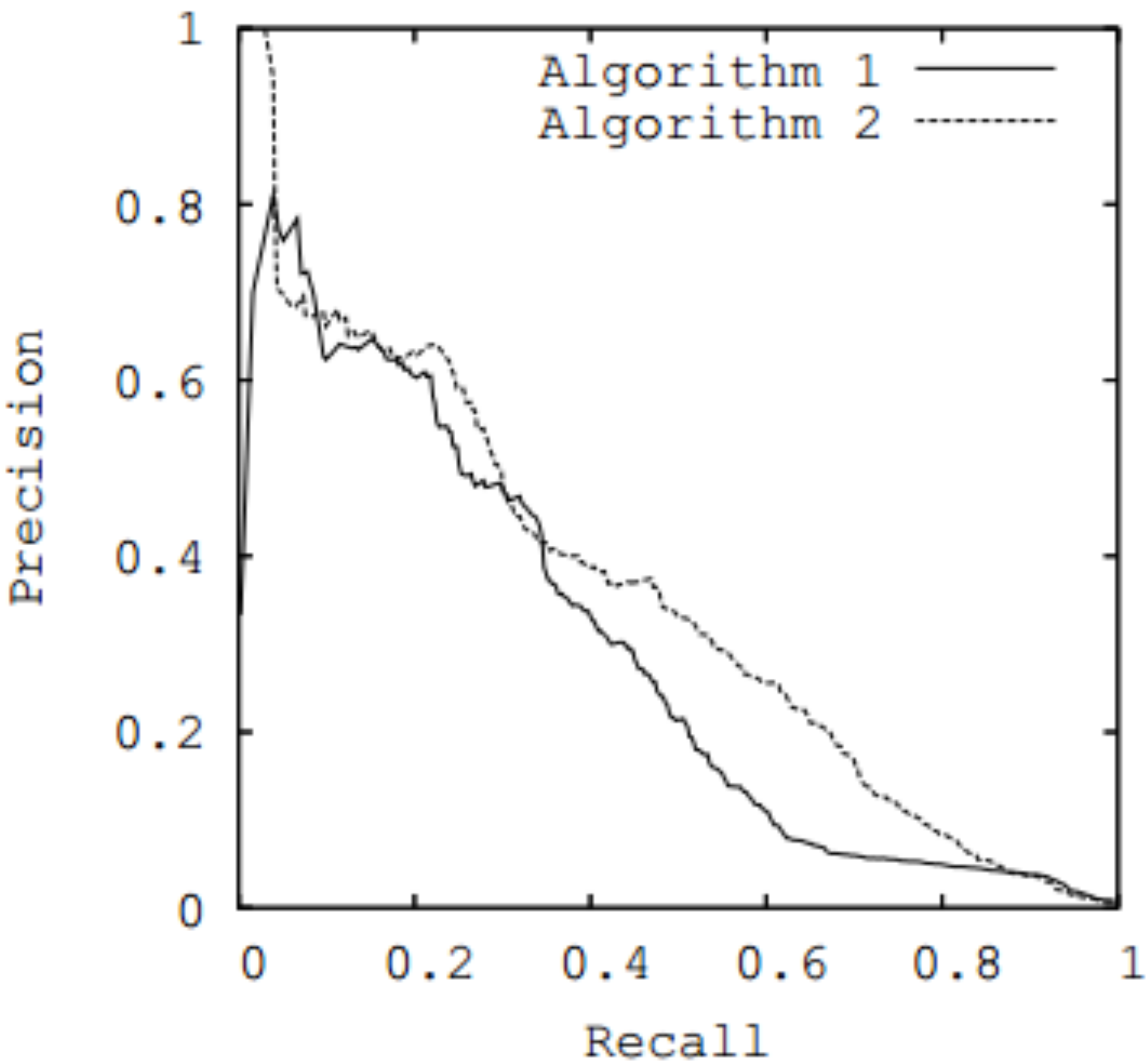


False positive rate, 1 - specificity

Average Precision for Imbalanced Classes or Retrieval Problem

- Suppose there are two classes:
  - Class 1 is much more prevalent than Class 2
  - What is a classifier with very high accuracy?
- Suppose we need to search for relevant documents from 1 billion webpages
  - Accuracy is not a good measurement

Precision-Recall Curve



Precision-Recall Curve

Average Precision (AP):

- Area under the Precision-Recall curve
- Summarize the whole curve

F1-score:

- Harmonic mean of a particular precision and recall

$$F_1 = \frac{2 * precistion * recall}{precision + recall}$$

Cross-Validation

- What if we have little data to split into separate disjoint train and test sets?
- Answer: Use cross-validation

K-fold Cross Validation

- Divide the data into K disjoint subsets
  - Train on the union of (K-1) subsets
  - Test on the left-out set
  - Repeat K-times, every subset is used for testing once.

Leave-one-out Cross Validation

- LOOCV is K-fold CV with K = N, the number of data points.

What You Need to Know

- Bias-Variance Tradeoff of Learner
- Train-error is NOT good estimate of Prediction-error
- Common error measurements:
  - Accuracy, Confusion Matrix
  - Precision-Recall, Average Precision
- Cross-Validation