CS578 Statistical Machine Learning Lecture 9

Jean Honorio
Purdue University

Goal of machine learning?

Goal of machine learning

- Use algorithms that will perform well in unseen data

Goal of machine learning

- Use algorithms that will perform well in unseen data

- How to measure performance?

- How to use unseen data?

Goal of machine learning

- Use algorithms that will perform well in unseen data

- How to measure performance?

- How to use unseen data?

- Variability?

- By-product: a way to set ***hyper-parameters***
  - e.g., *C* for SVMs, λ for kernel ridge regression, etc.

Measures of Performance: Regression

- Assume that for a point x , we predict $g(x)$
- Mean square error: $MSE(g) = \frac{1}{n}\sum_{i=1}^{n}(g(x_i) - y_i)^2$
- Root mean square error: $RMSE(g) = \sqrt{MSE(g)}$
- Mean absolute error: $\frac{1}{n}\sum_{i=1}^{n}\left|g(x_i) - y_i\right|$

Measures of Performance: Classification

- True Positive (*TP*)
- True Negative (*TN*)
- False Positive (*FP*)
- False Negative (*FN*)

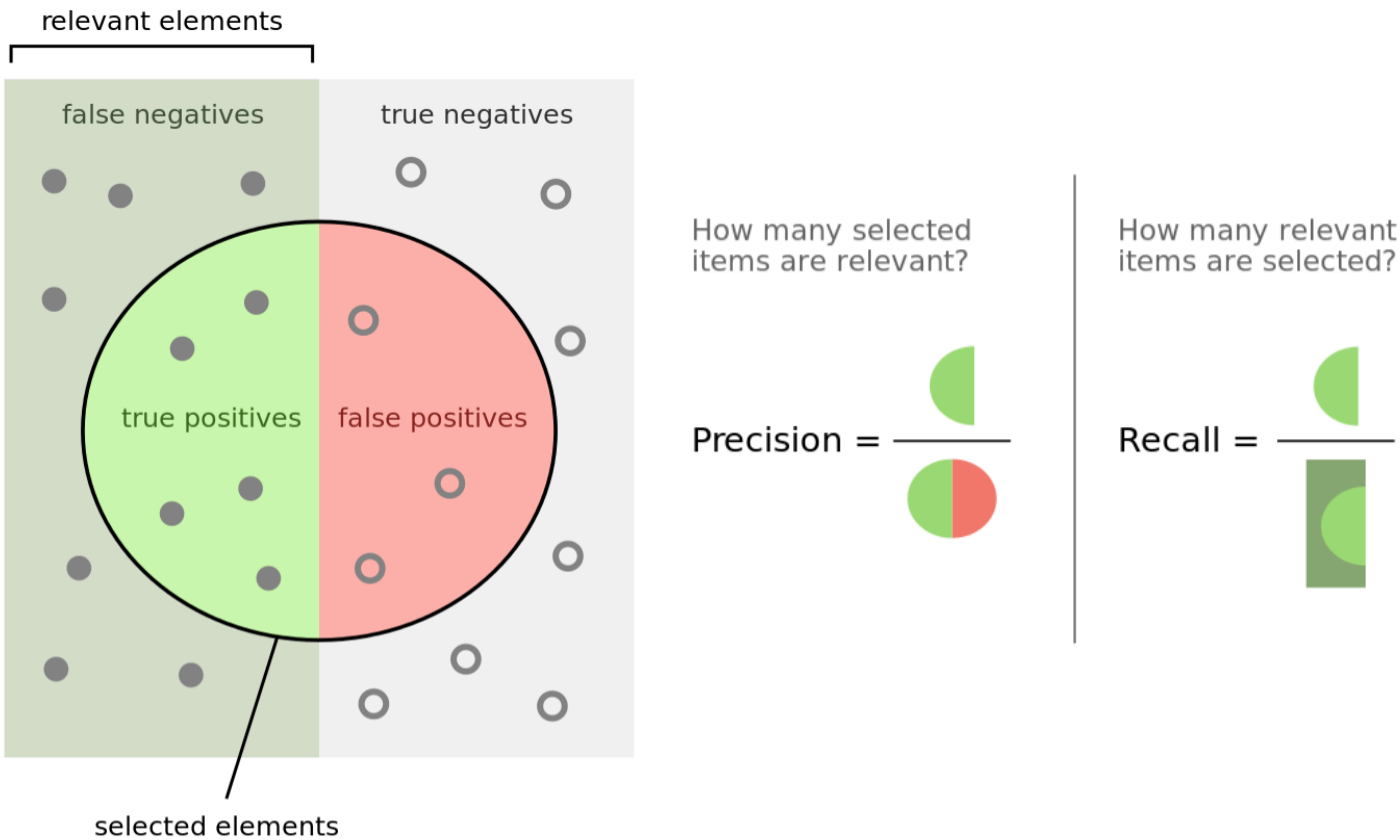
		True Label	
		+1	-1
Predicted Label	+1	*TP*	*FP*
	-1	*FN*	*TN*

- Accuracy $(TP + TN) / (TP + FP + FN + TN)$
- Error $(FP + FN) / (TP + FP + FN + TN)$
- Recall / Sensitivity $TP / (TP + FN)$
- Precision $TP / (TP + FP)$
- Specificity $TN / (TN + FP)$

- Use jointly: (Precision,Recall) or (Sensitivity,Specificity)

Precision and Recall

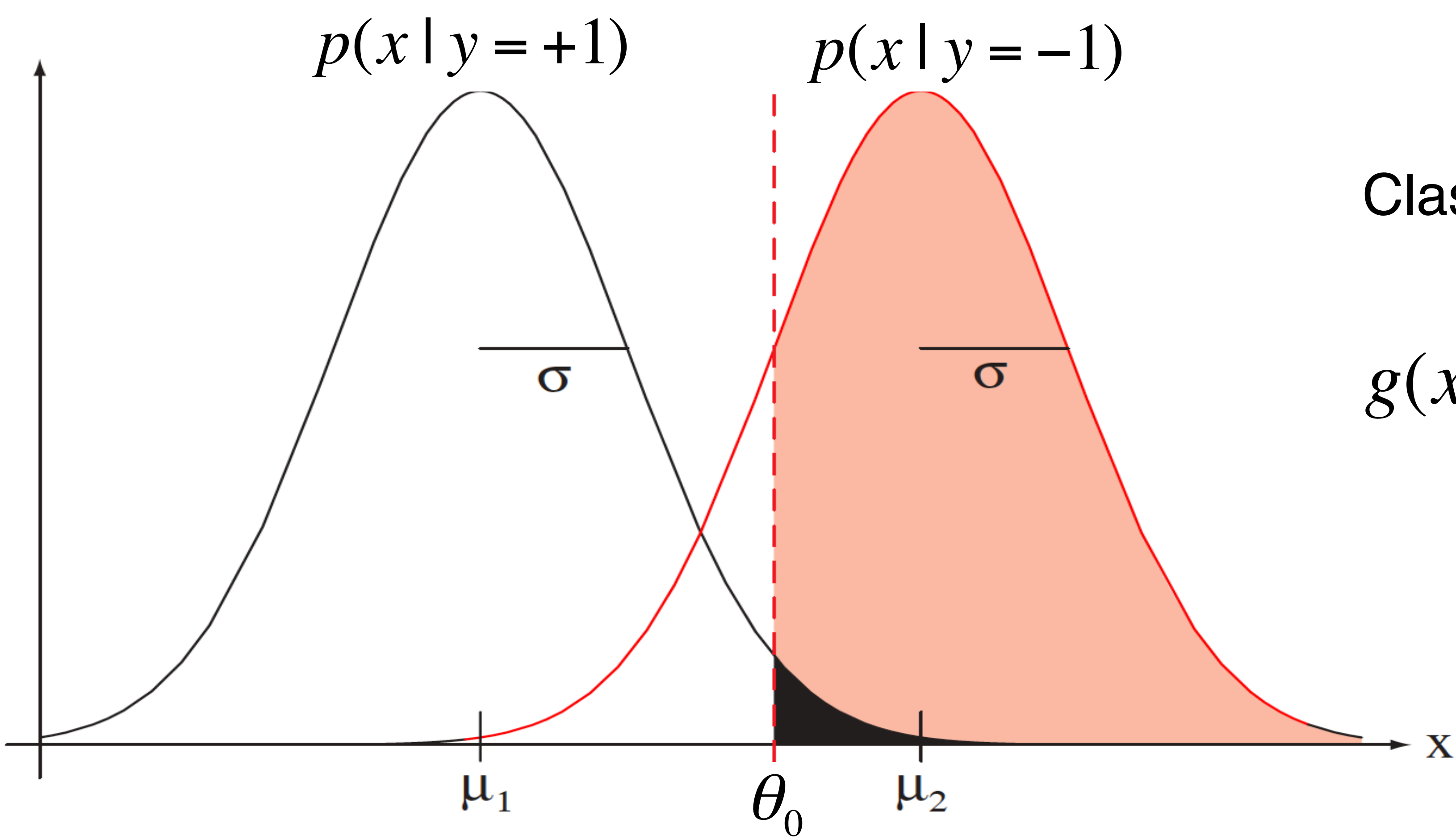
- Idea comes from information retrieval



Sensitivity and Specificity

- Idea comes from signal detection theory
- Assume Gaussian distributions $p(x \mid y=+1) = N(\mu_1, \sigma^2)$

$$p(x \mid y=-1) = N(\mu_2, \sigma^2)$$



Classifier:

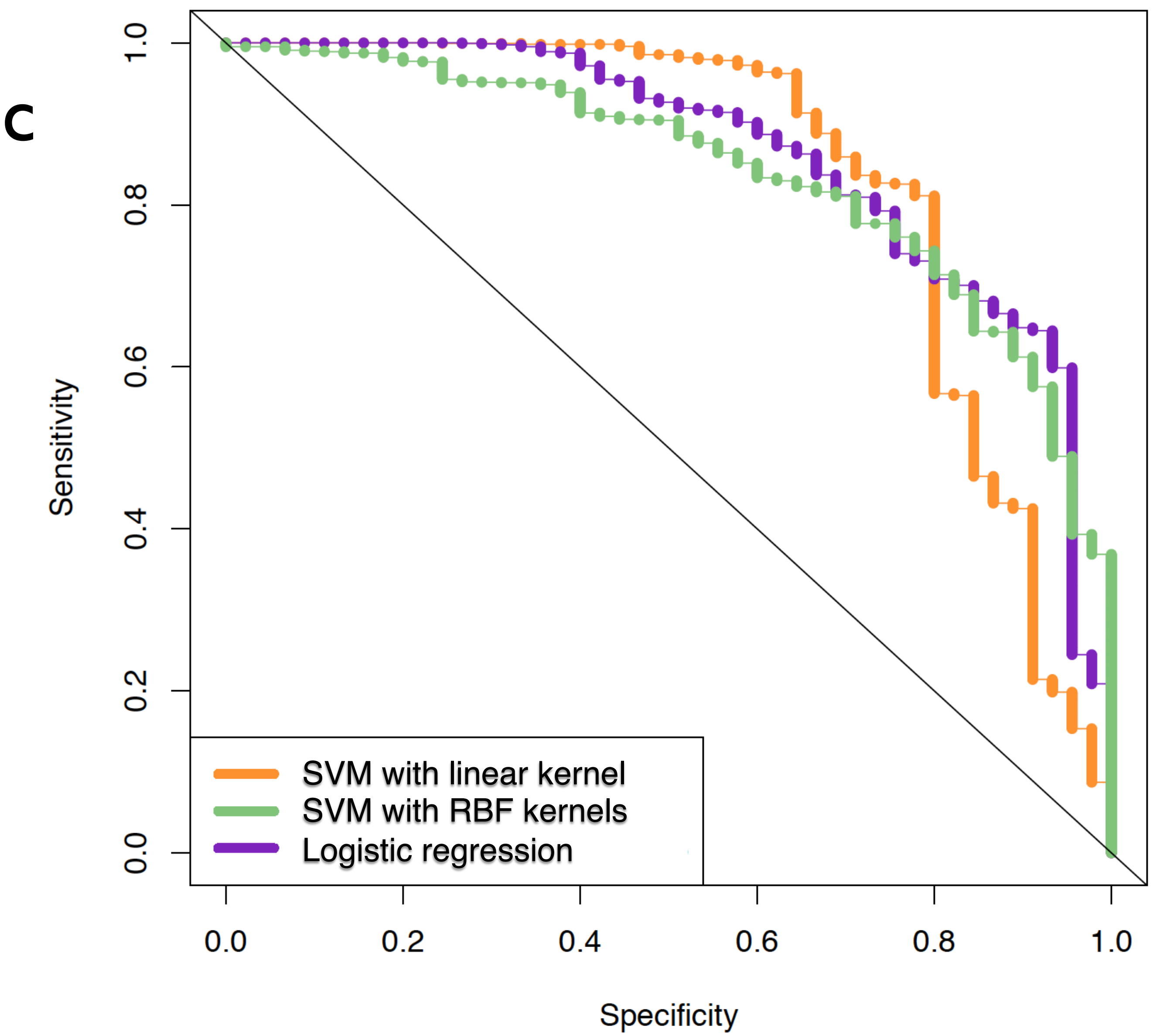
$$g(x) = \begin{cases} +1, & \text{if } x < \theta_0 \\ -1, & \text{if } x \geq \theta_0 \end{cases}$$

- By sliding the offset θ_0 we get different (*TP*, *FP*, *TN*, *FN*) and thus, different sensitivity and specificity

Receiver Operating Characteristic (ROC)

- By varying the offset for a classifier (e.g., SVMs, logistic regression) we can get different:
  - Sensitivity
  - Specificity

- Summarized with an Area Under the Curve (AUC)
  - Random: 0.5
  - Perfect classifier: 1



Other Loss Functions

- Let +1 mean “diseased patient” and -1 mean “healthy patient”

Predicted Label \ True Label	+1	-1
+1	0	1
-1	1	0

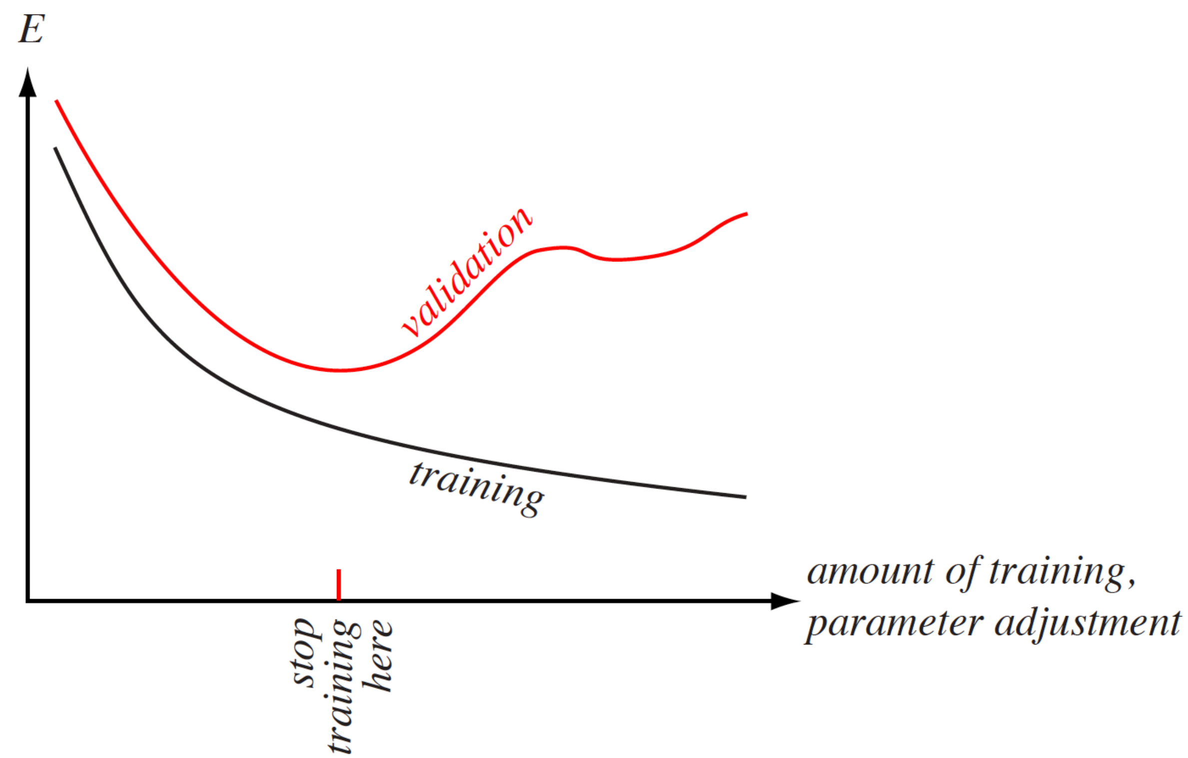
Predicted Label \ True Label	+1	-1
+1	0	1
-1	10	0

$$\frac{1}{n}\sum_{i=1}^{n} 1[g(x_i) \neq y_i]$$

$$\frac{1}{n}\sum_{i=1}^{n} Cost(g(x_i), y_i)$$

2) Using "Unseen" Data

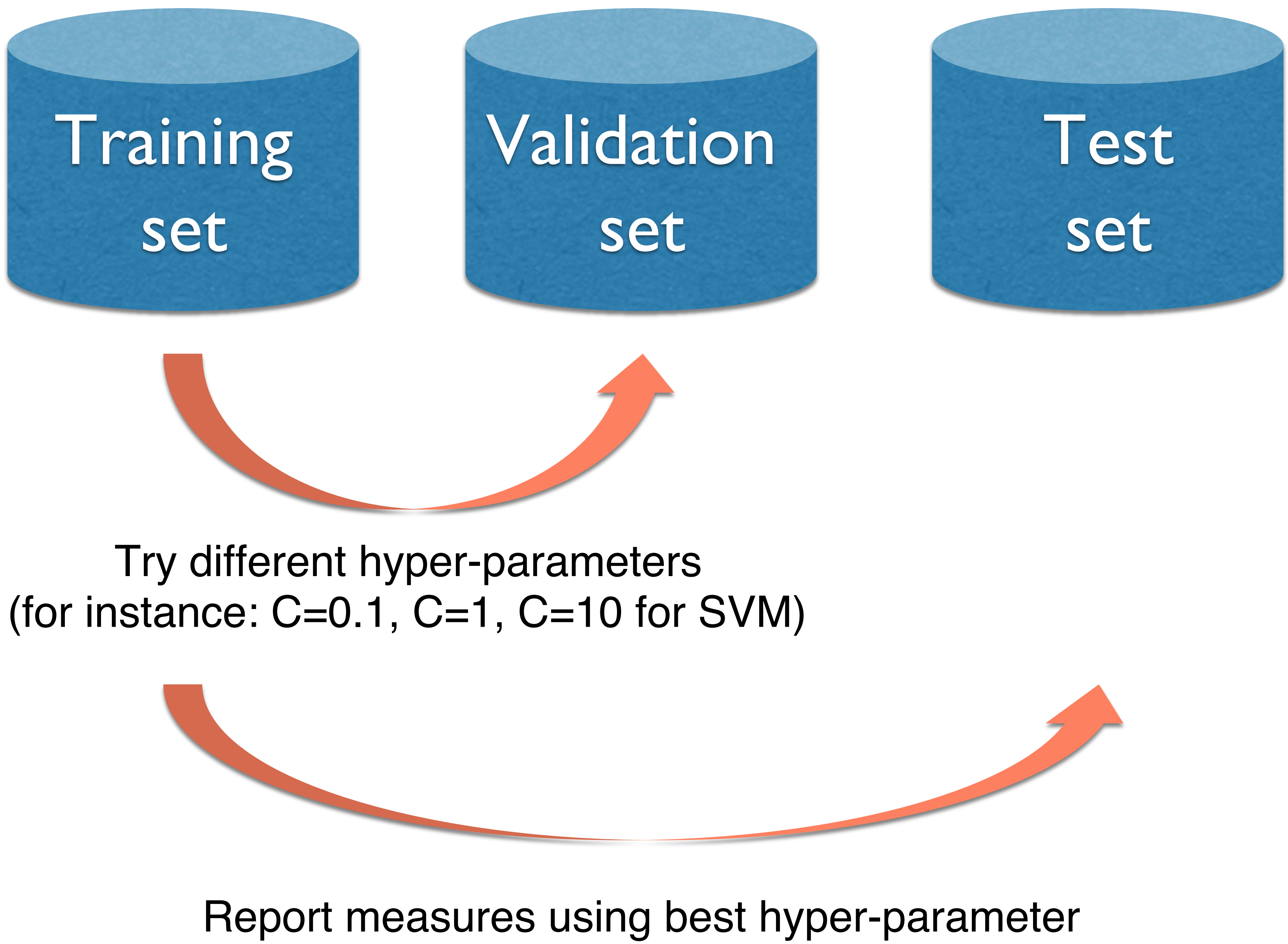
- Overfitting:
  - More complex methods fit better the training data (e.g., linear kernel versus cubic kernel)
  - Find hyper-parameters that better fit training data
  - Usually poor performance in unseen data



- To prevent overfitting, how can we "see" unseen data?
  - Simulate it !

Training, Validation, Testing

- Three data sets:



k-Fold Cross Validation

- Split training data D into k disjoint sets $S_1,\ldots,S_k$
  - Either randomly, or in a fixed fashion
  - If D has n samples, then each fold has approximately n / k samples
  - Popular choices: k=5, k=10, k=n (leave-one-out)

- For i = 1…k:

  train with sets $S_1,\ldots,S_{i-1}$, $S_{i+1},\ldots,S_k$

  test on set S_i

  let M_i be the test measure (e.g., accuracy, MSE)

- Mean and variance are:

$$\hat{\mu} = \frac{1}{k}\sum_{i=1}^{k} M_i \qquad\qquad \hat{\sigma}^2 = \frac{1}{k}\sum_{i=1}^{k} (M_i - \hat{\mu})^2$$

0.632 Bootstrapping

- Let B>0, and n be the number of training samples in D

- For i = 1…B:

  Pick n samples from D with replacement, call it S_i

  (S_i might contain the same sample more than once)

  train with set S_i

  test on the remaining samples ($D - S_i$)

  let M_i be the test measure (e.g., accuracy, MSE)

- Mean and variance are:

$$\hat{\mu} = \frac{1}{B}\sum_{i=1}^{B} M_i \qquad \hat{\sigma}^2 = \frac{1}{B}\sum_{i=1}^{B} (M_i - \hat{\mu})^2$$

0.632 Bootstrapping

- Why 0.632 ?

- Recall that:
  - We pick *n* items with replacement from out of *n* items
  - We choose uniformly at random

- The probability of:
  - not picking one particular item in 1 draw is $1-1/n$
  - not picking one particular item in *n* draws is $(1-1/n)^n$
  - picking one particular item in *n* draws is $1-(1-1/n)^n$

- Finally: $\lim_{n\to\infty} 1-\left(1-1/n\right)^n = 1-1/e \approx 0.632$

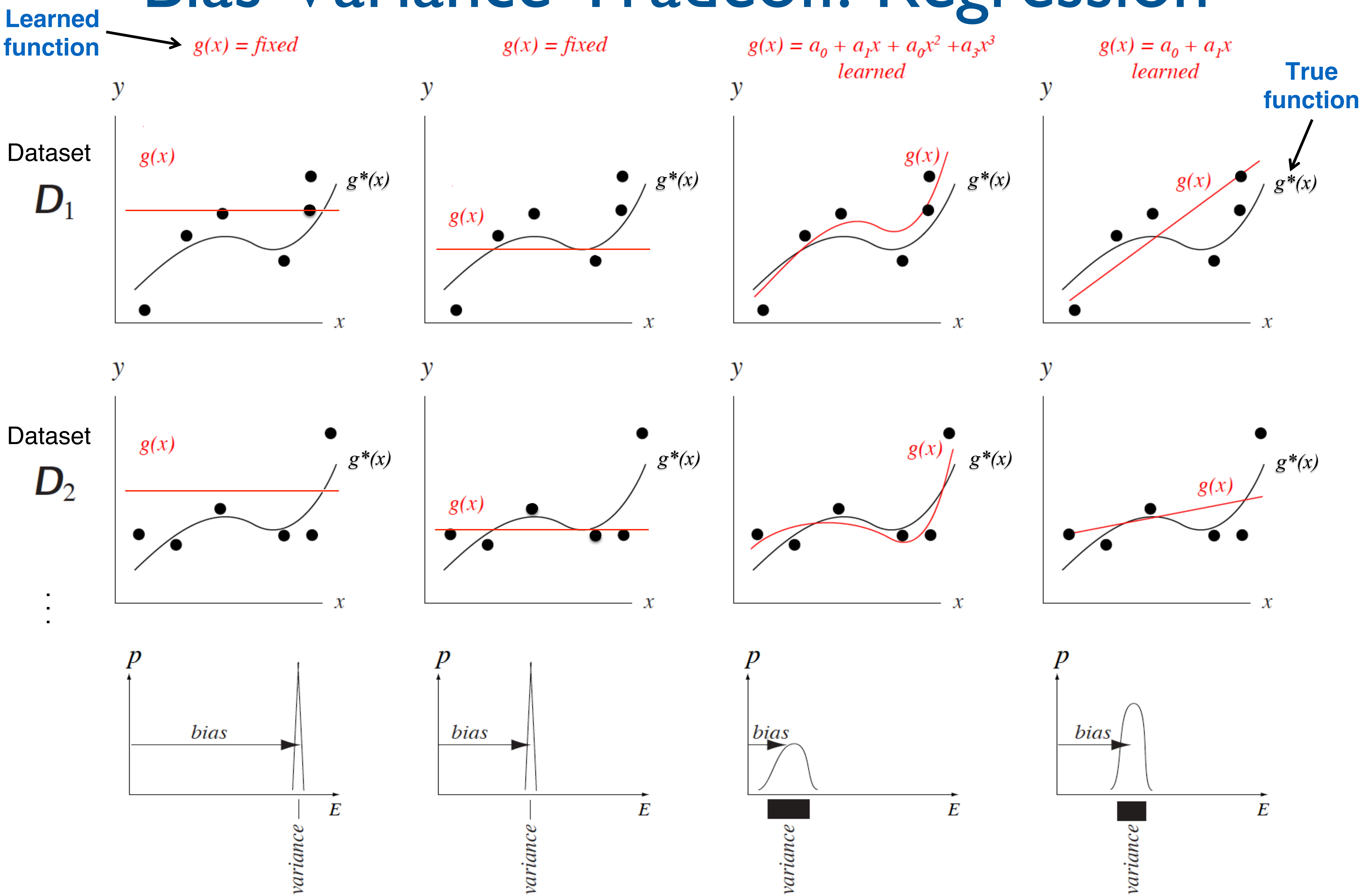
3) Variability

- How to compare two algorithms?
  - Not only means, also variances !

- Bias-variance tradeoff

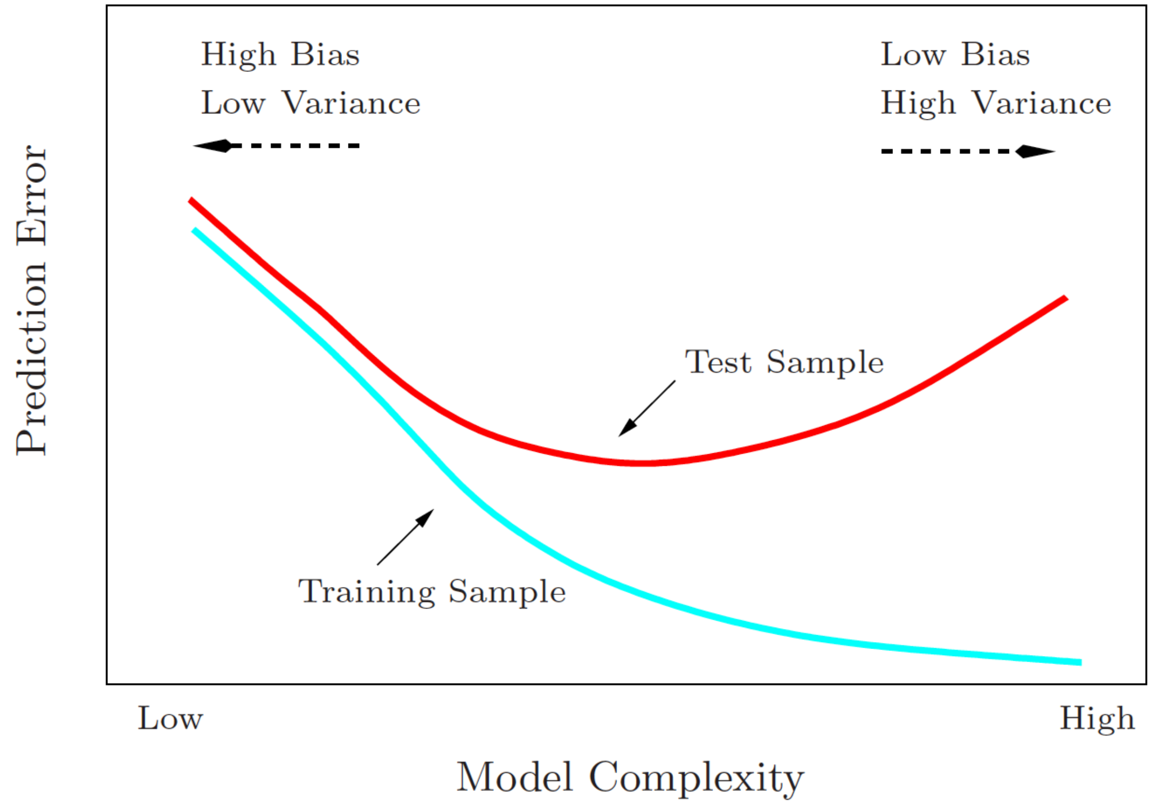
- Statistical hypothesis testing

Bias-Variance Tradeoff: Regression



Bias-Variance Tradeoff: Regression

- More complex methods (e.g., cubic kernel): low bias, high variance
- Less complex methods (e.g., linear kernel): high bias, low variance



- The mean squared error decomposes:

$$\mathcal{E}_{\mathcal{D}}\left[(g(\mathbf{x};\ \mathcal{D}) - g^*(\mathbf{x}))^2\right]$$
$$= \underbrace{(\mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D}) - g^*(\mathbf{x})])^2}_{bias^2} + \underbrace{\mathcal{E}_{\mathcal{D}}\left[(g(\mathbf{x};\ \mathcal{D}) - \mathcal{E}_{\mathcal{D}}[g(\mathbf{x};\ \mathcal{D})])^2\right]}_{variance}$$

Statistical Hypothesis Testing

- How to compare two algorithms?
  - Not only means, also variances !

- Let $\hat{\mu}_1, \hat{\sigma}_1^2, \hat{\mu}_2, \hat{\sigma}_2^2$ be mean and variance of algorithms 1 and 2.

- When to reject null hypothesis $\mu_1 = \mu_2$ in favor of $\mu_1 > \mu_2$?

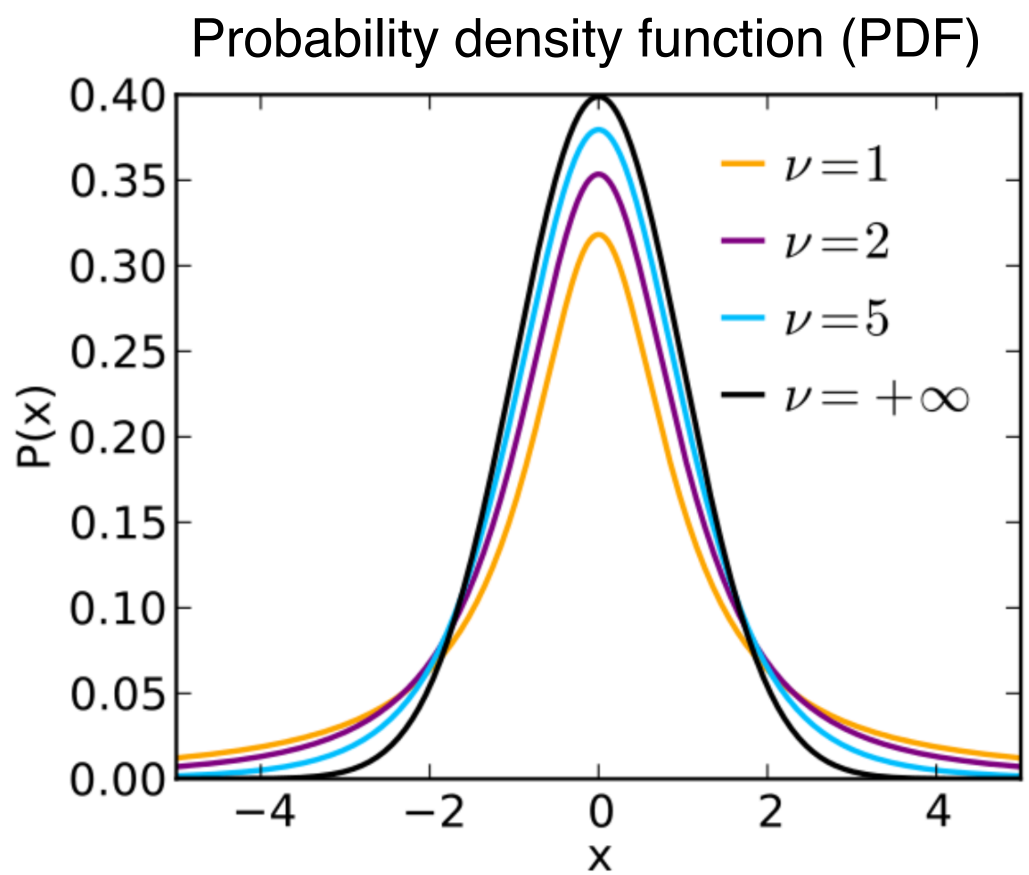
- Let:

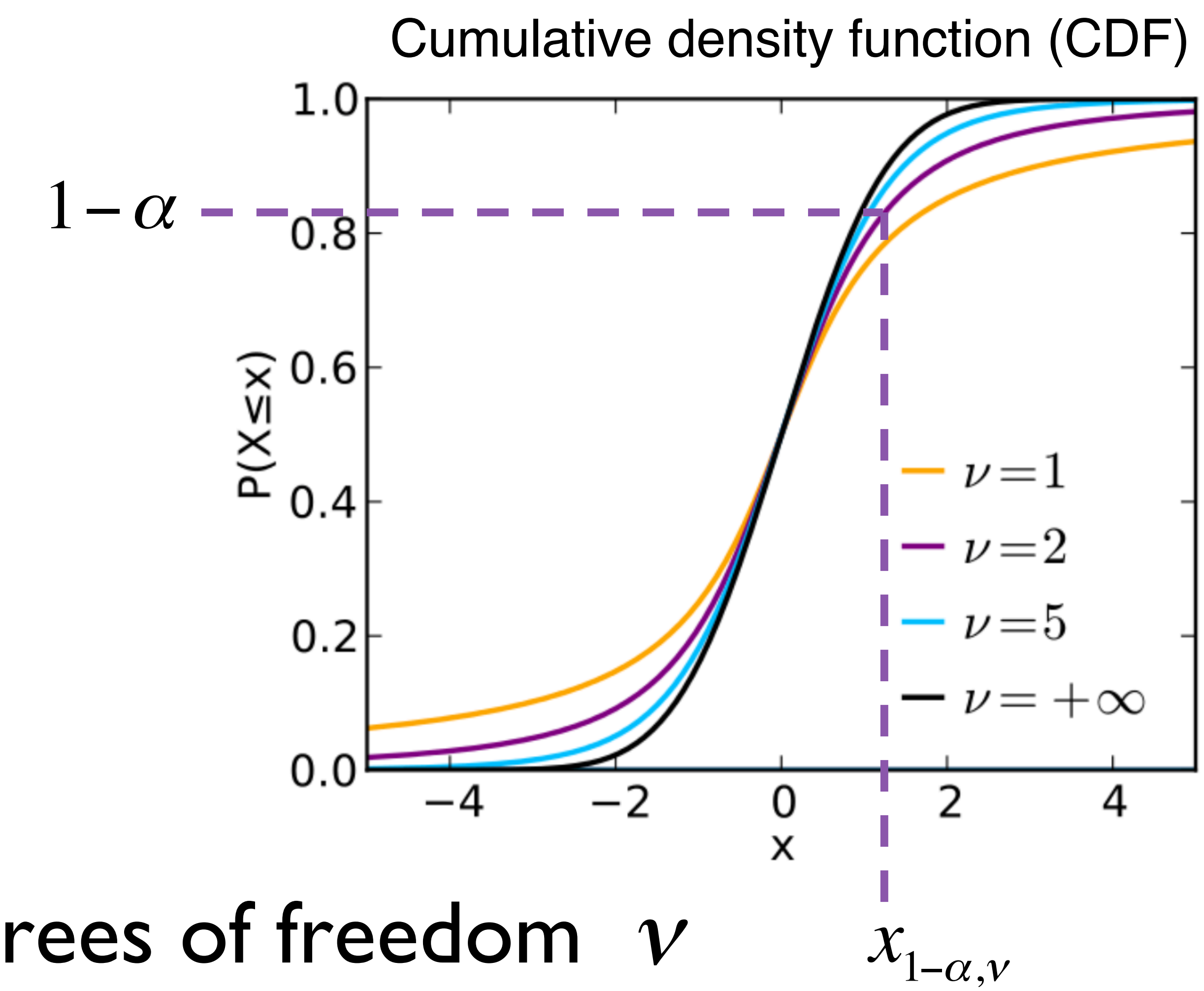
$$x = \frac{(\hat{\mu}_1 - \hat{\mu}_2)\sqrt{n}}{\sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}} \qquad \nu = \left\lceil \frac{(\hat{\sigma}_1^2 + \hat{\sigma}_2^2)^2 (n-1)}{\hat{\sigma}_1^4 + \hat{\sigma}_2^4} \right\rceil$$

Degrees of freedom of Student's t-distribution

Statistical Hypothesis Testing

- Student's t-distribution:





- For significance level α, degrees of freedom ν
  - Find the value $x_{1-\alpha,\nu}$ for which CDF = $1-\alpha$
  - Matlab: tinv(1–alpha, v)
- If $x > x_{1-\alpha,\nu}$ reject null hypothesis $\mu_1 = \mu_2$ in favor of $\mu_1 > \mu_2$

Statistical Hypothesis Testing: Example I

- Two algorithms tested with 9-fold cross validation
- Percentage of error on each left-out fold:
  - A1: 11, 7, 13, 12, 12, 9, 10, 7, 10 $\qquad \hat{\mu}_1 = 10.1, \quad \hat{\sigma}_1^2 = 4.1$
  - A2: 10, 8, 12, 10, 11, 9, 13, 7, 9 $\qquad \hat{\mu}_2 = 9.9, \quad \hat{\sigma}_2^2 = 3.2$

- Can we reject null hypothesis ($\mu_1 = \mu_2$) in favor of alternate hypothesis ($\mu_1 > \mu_2$) at **5%** significance level?

$$x = \frac{(\hat{\mu}_1 - \hat{\mu}_2)\sqrt{n}}{\sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}} = \frac{(10.1 - 9.9)\sqrt{9}}{\sqrt{4.1 + 3.2}} \approx \frac{0.2 \times 3}{2.7} \approx 0.22$$

$$\nu = \left\lceil \frac{(\hat{\sigma}_1^2 + \hat{\sigma}_2^2)^2 (n-1)}{\hat{\sigma}_1^4 + \hat{\sigma}_2^4} \right\rceil = \left\lceil \frac{(4.1 + 3.2)^2 (9-1)}{4.1^2 + 3.2^2} \right\rceil \approx \left\lceil \frac{7.3^2 \times 8}{27} \right\rceil \approx \lceil 15.8 \rceil = 16$$

- Inverse CDF $x_{1-0.05,\nu} = x_{0.95,16} = 1.75$

$x = 0.22 \le 1.75 = x_{0.95,16}$ then ***cannot reject null***

Statistical Hypothesis Testing: Example 2

- Two algorithms tested with 9-fold cross validation
- Percentage of error on each left-out fold:
  - A1: 10, 12, 14, 13, 13, 10, 11, 10, 11 $\quad \hat{\mu}_1 = 11.6, \quad \hat{\sigma}_1^2 = 2$
  - A2: 10, 8, 12, 10, 11, 9, 13, 7, 9 $\quad \hat{\mu}_2 = 9.9, \quad \hat{\sigma}_2^2 = 3.2$

- Can we reject null hypothesis ($\mu_1 = \mu_2$) in favor of alternate hypothesis ($\mu_1 > \mu_2$) at **5%** significance level?

$$x = \frac{(\hat{\mu}_1 - \hat{\mu}_2)\sqrt{n}}{\sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}} = \frac{(11.6 - 9.9)\sqrt{9}}{\sqrt{2 + 3.2}} \approx \frac{1.7 \times 3}{2.3} \approx 2.22$$

$$\nu = \left\lceil \frac{(\hat{\sigma}_1^2 + \hat{\sigma}_2^2)^2 (n-1)}{\hat{\sigma}_1^4 + \hat{\sigma}_2^4} \right\rceil = \left\lceil \frac{(2 + 3.2)^2 (9-1)}{2^2 + 3.2^2} \right\rceil \approx \left\lceil \frac{5.4^2 \times 8}{14.2} \right\rceil \approx \lceil 16.5 \rceil = 17$$

- Inverse CDF $x_{1-0.05,\nu} = x_{0.95,17} = 1.74$

$x = 2.22 > 1.74 = x_{0.95,17}$ then ***reject null***

What is a Sample?

- In this lecture we assume that each sample is a different "unit of interest" for the experimenter
- Never sample the same "unit of interest" several times
  - In a medical application, we might be interested on knowing the accuracy (and variance) with respect to patients.
  - Taking two visits of the same patient as two different samples would be incorrect.
- Collect more data, if necessary
  - Never duplicate (copy & paste) data.