



# EE 6885 Statistical Pattern Recognition

Fall 2005  
Prof. Shih-Fu Chang  
<http://www.ee.columbia.edu/~sfchang>

Lecture 16 (11/21/05)

EE6887-Chang

16-1

- Today's lecture
  - SVM tool using libsvm
  - Analysis of Classification Algorithms
    - Bias vs. Variance, DHS Chap 9.3
    - Bagging, Boosting, DHS Chap 9.5
- Homework #7 due Nov. 30th
- Final Exam (note: date change from CU schedule)
  - Dec. 16<sup>th</sup> Friday 1:10-4 pm, Mudd Rm 644

EE6887-Chang

16-2

## Project: Image Classification

- Topics: feature selection, model fusion, SVM optimization
- Data set: TRECVID 05 images
  - 3504 images
  - 555 buildings, 136 explosion\_fire
    - Note the labels may not be correct.  
Don't correct them. Treat them as noisy labels.
- Features:
  - Grid color moments (225 dim) and Gabor texture (48 dim)
- Requirements: 5-10-page report and slides, Due Dec. 12
- Focus:
  - Get familiar with baseline approaches
  - Pursue specific issues for deep investigation!
- What classifiers and features to use?



EE6887-Chang

16-3

## Selection of the right learning algorithm

- "No Free Lunch Theorem"
  - No learning or classification method is superior to others universally if we don't have prior information.
    - Need to consider problem context and intended usage.
  - Given a problem, we should consider
    - Prior information (domain knowledge)
    - Data distributions
    - Amount of training data
    - Cost of reward functions
- Occam's Razor principle
  - one should not use classifiers that are more complicated than are necessary.

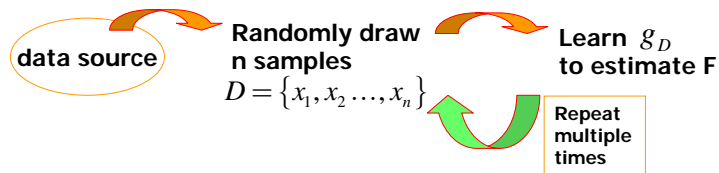
EE6887-Chang

16-4

## Bias and Variance of Estimators

- Validation paradigm
  - train classifiers using some training data  $D$
  - measure the generalization performance of a classifier over test data
  - conduct K-fold cross validation
- We want to know how the generalization performance varies when the training data changes

Assume  $F$  is a quantity whose value is to be estimated



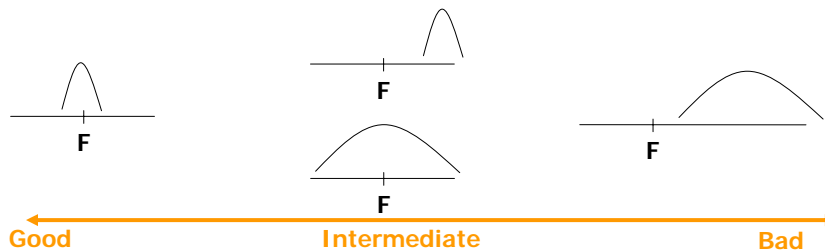
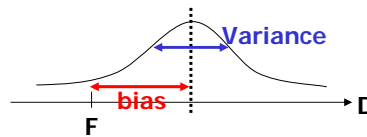
expected error:  $E_D \left[ |g_D - F|^2 \right]$

EE6887-Chang

16-5

## Bias and variance (2)

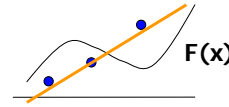
$$\begin{aligned}
 \text{expected error: } & E_D \left[ |g_D - F|^2 \right] \\
 &= E_D \left[ g_D^2 - 2g_DF + F^2 \right] \\
 &= \underbrace{\left[ E_D(g_D) - F \right]^2}_{\text{Bias}^2} + \underbrace{E_D \left[ |g_D - E_D(g_D)|^2 \right]}_{\text{Variance}}
 \end{aligned}$$



EE6887-Chang

16-6

# Bias and variance of Regression



Want to estimate a function  $F(x)$

From training data  $D = \{x_1, F(x_1), x_2, F(x_2), \dots, x_n, F(x_n)\}$

Learn a regression function  $g(x; D)$

For a given  $D$ , estimation error:  $\int_x |g(x; D) - F(x)|^2 p(x) dx$

$$\begin{aligned} \text{Expected estimation error: } & E_D \left[ \int_x |g(x; D) - F(x)|^2 p(x) dx \right] \\ &= \int_x E_D \left[ |g(x; D) - F(x)|^2 \right] p(x) dx \\ &= \int_x \left( \text{Bias}(x)^2 + \text{Variance}(x) \right) p(x) dx \\ &= \text{Bias}^2 + \text{Variance} \end{aligned}$$

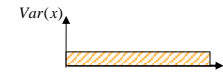
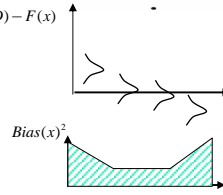
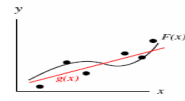
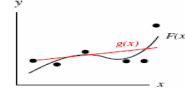
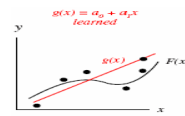
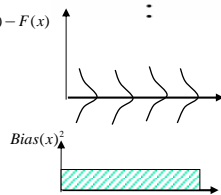
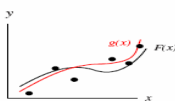
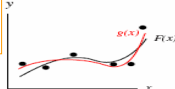
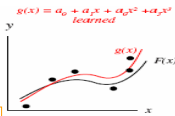
EE6887-Chang

16-7

## Example (cubic source)

$$g(x) = a_0 + a_1x + a_2x^2 + a_3x^3$$

(learned)



$$g(x) = a_0 + a_1x$$

(learned)

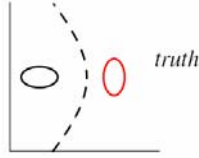
■ How do Bias and Variance change when there are more training data?

EE6887-Chang

16-8

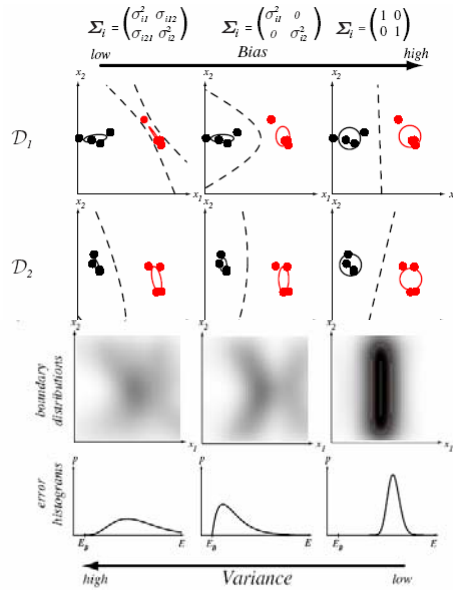
## Bias vs. variance for classification

- Ground truth: 2D Gaussian



- Complex models have more variances than simple models
- Increasing training pool size helps reduce the variance
- Decision boundaries more stable when models are simpler
- Prior knowledge helps! (see the middle column)

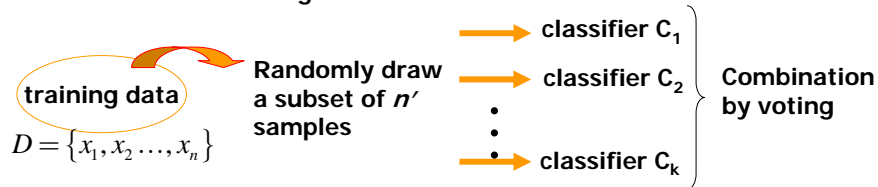
EE6887-Chang



16-9

## Bagging

- Classifiers are affected by the choices of training set
  - train multiple component classifiers by creating multiple subsets of training data



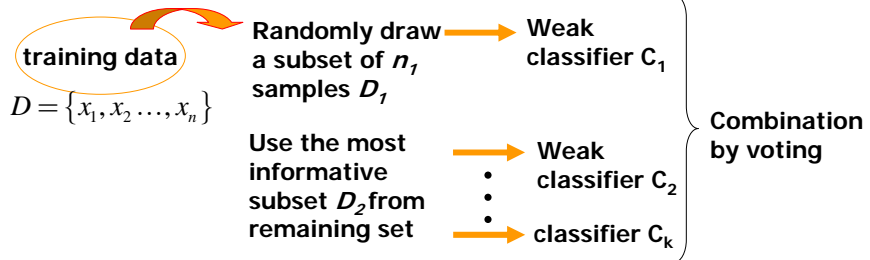
- When does this help?
- Issues
  - $n'$
  - sampling scheme
  - number of component classifiers
  - fusing method

EE6887-Chang

16-10

# Boosting

- For each component classifier, use the subset of data that is most informative given the current set of component classifiers

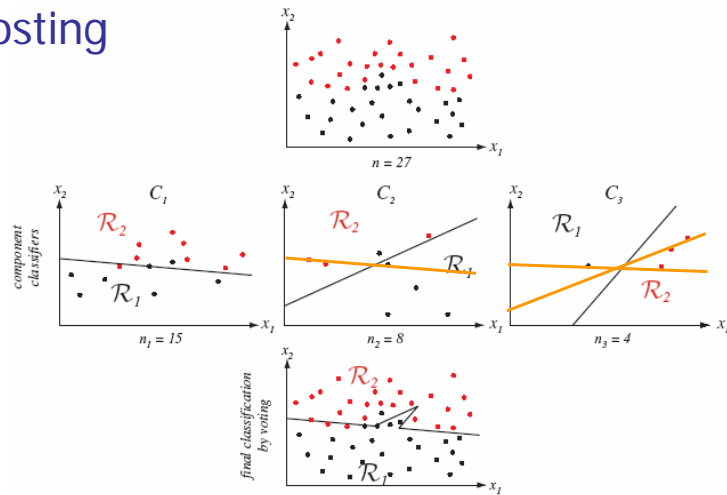


- What's the most informative subset for  $C_2$ ?
  - only data that are misclassified by  $C_1$ ?
  - (OR) half correctly classified and half misclassified?
- What's the most informative subset for  $C_3$ ?
  - include only points that  $C_1$  and  $C_2$  cannot agree, use  $C_3$  as tie breaker

EE6887-Chang

16-11

# Boosting



EE6887-Chang

16-12

## AdaBoost

- Add weak classifiers until low training error has been achieved
- Each training pattern receives a weight determining its chance of being selected for subsequent learning steps.
- If a pattern is correctly classified, then the weight is decreased.

begin with  $W(i) = 1/n, i = 1, \dots, n$

$k = 1, \dots, k_{\max}$

train weak classifier  $C_k$  using  $D$  training patterns with weights  $W_k(i)$

$E_k \leftarrow$  training error of  $C_k$  measured on  $D$  using weights  $W_k(i)$

$$\alpha_k \leftarrow \frac{1}{2} \ln[(1 - E_k) / E_k]$$

$$W_{k+1}(i) \leftarrow \frac{W_k(i)}{Z_k} \times \begin{cases} e^{-\alpha_k}, & \text{if } x_i \text{ is correctly classified} \\ e^{\alpha_k}, & \text{if } x_i \text{ is incorrectly classified} \end{cases}$$

final classification

$$g(\mathbf{x}) = \sum_{k=1}^{k_{\max}} \alpha_k h_k(\mathbf{x}), \text{ where } h_k(\mathbf{x}) \text{ is the predicted output from } C_k$$

EE6887-Chang

16-13

## AdaBoost



ensemble training error rate

$$E = \prod_{k=1}^{k_{\max}} \left[ 2\sqrt{E_k(1 - E_k)} \right]$$

EE6887-Chang

16-14