Statistical Learning Machines (SLM): Design, Assessment, and Advice for Practitioners

Faculty of Computers and Information

Waleed A. Yousef

Computer Science Department,
Faculty of Computers and Information,
Helwan University

10^0

meter

Distance to a bunch of leaves

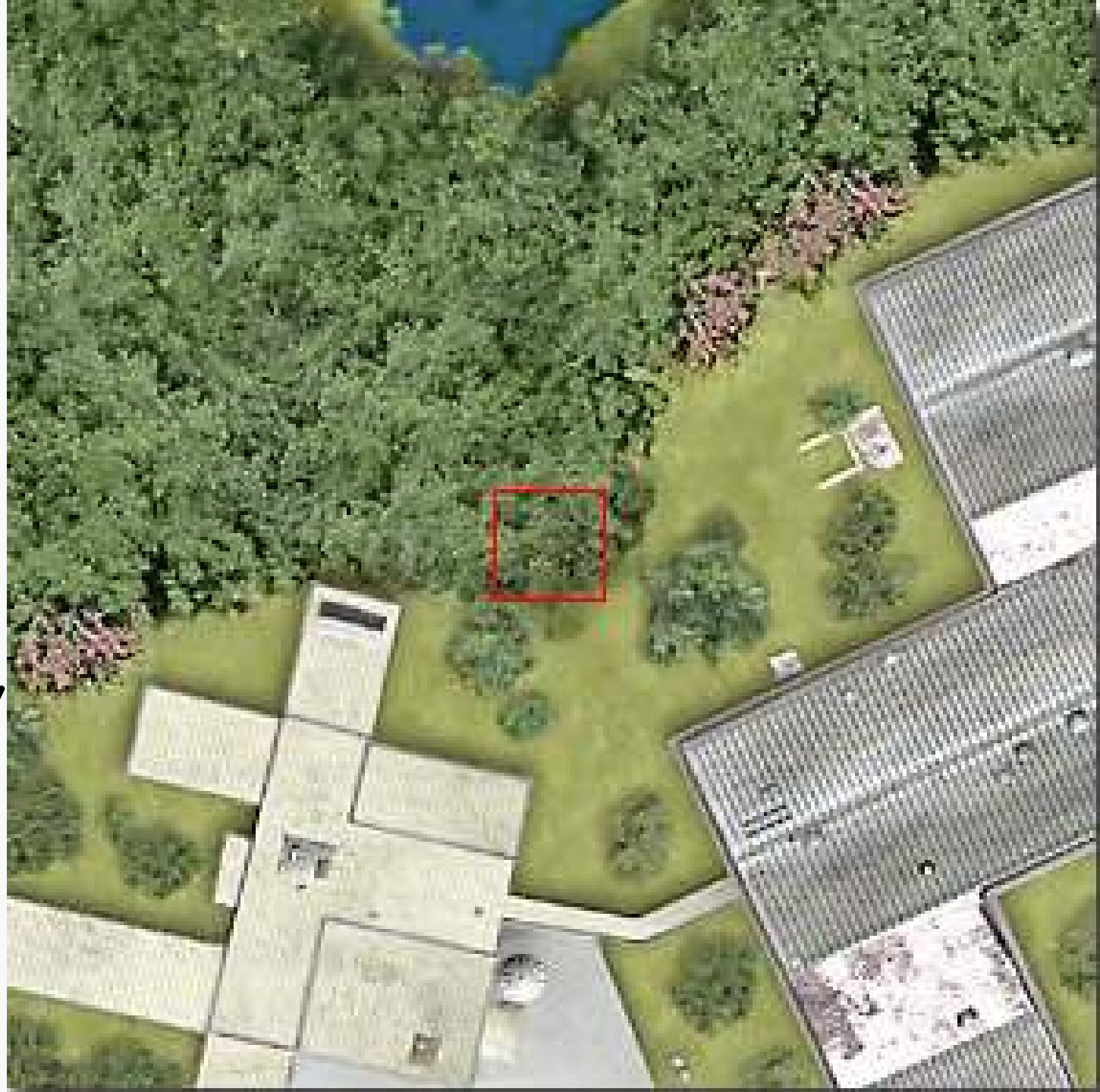
10^1

meter

Forest?!

10^2 meter

Now,
you can see your whole landscape;
you can easily connect pieces together if you see the big picture.



Objectives & Audience

- See the whole picture of the SLM for better understanding and insight.
- Emphasize the importance of mathematical rigor; particularly Statistics and Probability
- Allude to some fallacies, pitfalls, and misconceptions in the field.

- **Practitioners**: for guidance and seeing the whole roadmap before starting their applications to avoid pitfalls.
- **Rigorous students**: in the middle of the "*Pattern Recognition*" course for seeing the whole picture again, summarizing the first half of the course and introducing the rest.

Contents

- Fundamental framework; Statistical Decision Theory (SDT):
  - SDT; case of regression.
  - SDT; case of classification.
  - Statistical Learning from data, and two subfields.

- Design:
  - Different methods for regression in one picture.
  - Different methods for classification in one picture.

- Assessment:
  - Model Complexity and Bias-Variance tradeoff.
  - Cover's Theorem
  - Different measures.
  - Different paradigms.
  - Different estimators.

- Concluding Remarks

Fundamental framework: SDT

$Y, X \sim r.v.$ with joint pdf f_{XY}.

Given $X = x$, what is the best guess of Y (call it $\hat{Y} = \eta^*(X)$)?

Best in the sense that the Expected loss $\mathrm{E}\, L(Y, \hat{Y})$ is minimized.

- The most common L when Y is quantitative is the square-loss:
  $L(Y, \hat{Y}) = (Y - \hat{Y})^2$; this is the case of regression.
- The most common L when Y is qualitative, i.e., $Y = \omega_1, \omega_2, \ldots, \omega_k$, is:
  $L(\omega_i, \omega_j)$; this is the case of classification.

SDT: case of regression

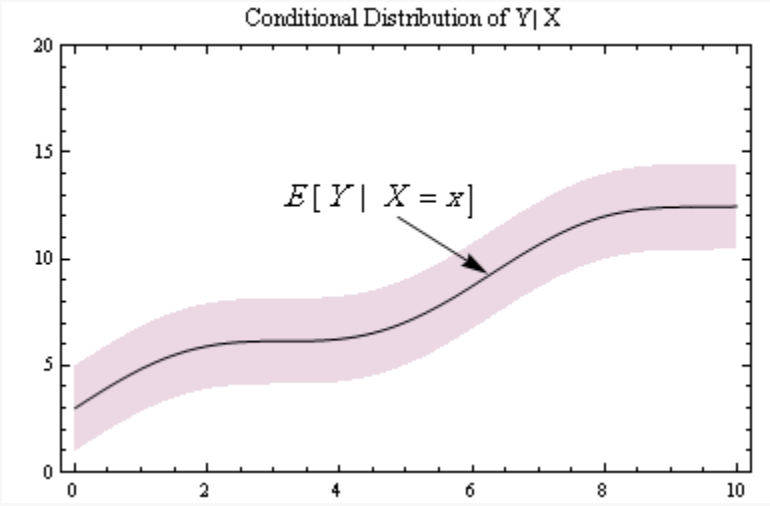
$$Y \in \mathbb{R}$$

$L(Y, \hat{Y}) = (Y - \hat{Y})^2$, then

$$\eta^*(X) = \arg\min \mathrm{E}(Y - \eta(X))^2$$

$$\vdots$$

$$= \mathrm{E}_Y(Y \mid X)$$



The minimum risk is given by :

$$R_{\min}(\eta^*) = E_X\left[\mathrm{var}_{Y|X}[Y \mid X]\right]$$

The risk, under square-loss function, is minimized iff

$\hat{Y} = \mathrm{E}_Y(Y \mid X)$; so, it is unique!

SDT; case of classification

$$Y \in \{\omega_1, \omega_2, \ldots, \omega_k\}$$

$$L(Y, \hat{Y}) = L(\omega_i, \omega_j) = c_{ij}, \text{ then}$$

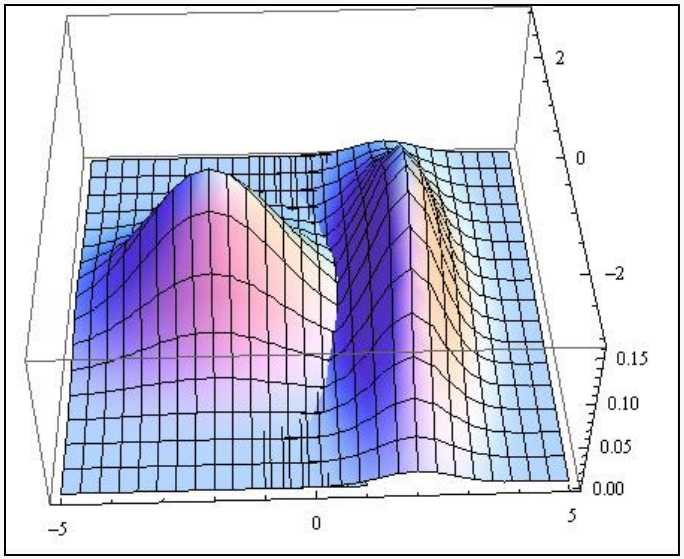
$$\eta^*(X) = \arg\min \mathrm{E}\, c_{ij}$$

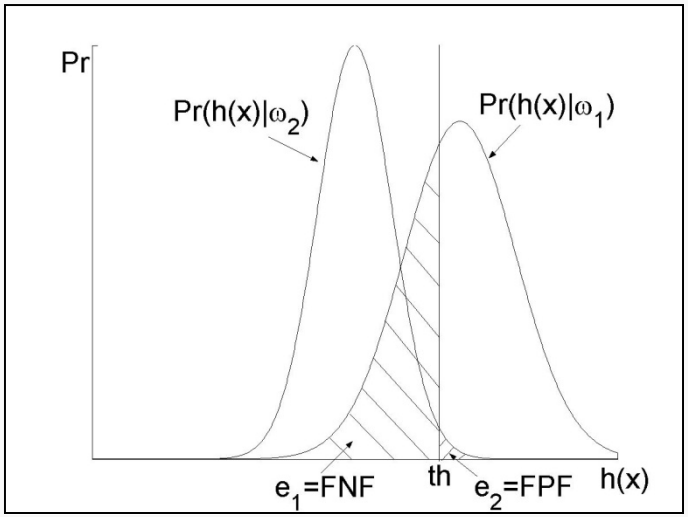
$$\vdots$$

$$= \arg\min_j \sum_{i=1}^{k} c_{ij} \Pr(Y = \omega_i \mid X)] \quad (*)$$

For the case of $k = 2$, and $c_{ii} = 0$

$$\frac{f_X(X = x \mid \omega_1)}{f_X(X = x \mid \omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{\Pr(\omega_2)(\, c_{21}\,)}{\Pr(\omega_1)(\, c_{12}\,)}$$





$$h(X) \underset{\omega_2}{\overset{\omega_1}{\gtrless}} th, \text{ where}$$

$$h(X) = \log \frac{f_X(X = x \mid \omega_1)}{f_X(X = x \mid \omega_2)}$$

$$th = \log \frac{\Pr(\omega_2)(\, c_{21}\,)}{\Pr(\omega_1)(\, c_{12}\,)}$$

The risk is minimized iff

$(*)$ is satisfied; so, it is unique!

Bayes classifier has min. Bayes risk:

$$R(\eta^*) = FPF\, c_{12}\, P_1 + FNF\, c_{21}\, P_2$$

$$Err(\eta^*) = FPF\, P_1 + FNF\, P_2$$

Avoid the Fallacy:

Some practitioners think of some decision rules (e.g., neural networks) as superior to all other rules.

The only one with this superiority is the Bayes' rule, which is unique.

" There has been a great deal of hype surrounding neural networks, making them seem magical and mysterious. …" [ESL]

[ESL] *Hastie, T., R. Tibshirani, and J.H. Friedman (2001), The elements of statistical learning: data mining, inference, and prediction. Springer series in statistics., New York: Springer*

Statistical Learning (SL) from data, and two subfields

***Definition**: Learning is the process of estimating an <u>unknown</u> input-output dependency or structure of a system using a <u>limited number of observations,</u>* $\mathbf{tr}$

Again: this means that the for the estimated rule $\eta_{\mathbf{tr}}(X)$

$$\eta_{\mathbf{tr}}(X) \neq \eta^*(X) \text{ and } R(\eta_{\mathbf{tr}}) > R(\eta^*).$$

Design

How to construct $\eta_{\mathbf{tr}}(X)$*?*

Assessment

How to assess $\eta_{\mathbf{tr}}(X)$*?*

Estimate the performance, e.g., $\hat{R}(\eta_{\mathbf{tr}})$

Estimate the uncertainty in your estimator $\widehat{\mathrm{var}}\hat{R}(\eta_{\mathbf{tr}})$

Formalizing SL

- The training set $\mathbf{tr} = \{t_i, i = 1, \ldots, n_{tr}\}$ has size n_{tr}
- An observation $t_i = (x_i, y_i) \in \mathbf{tr}$, x_i is predictor and y_i is response.
- y_i is known; this is called Supervised Learning.
- Design (construct) a decision rule $\eta_{\mathbf{tr}}$ such that for $t_0 \notin \mathbf{tr}$

  $\eta_{\mathbf{tr}}(x_0)$ is close to y_0 in "some sense", e.g., $R(\eta_{\mathbf{tr}})$

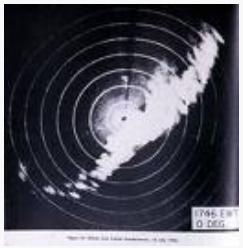
  That is, minimizing $\mathrm{E}\, L(y_0, \eta_{\mathbf{tr}}(x_0))$

Why Learning?

- Understanding
- Prediction

SLM and Other Fields of Science

Applications | **Other Fields** | **SLM**

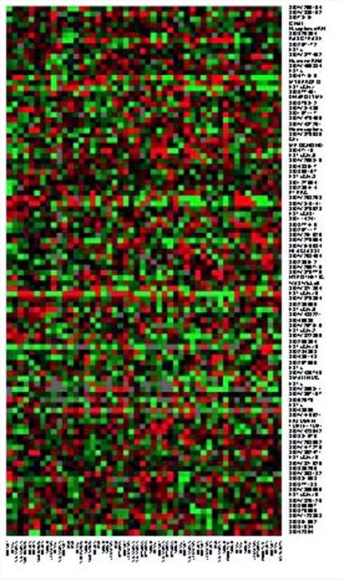


Automatic Target Recognition (ATR)

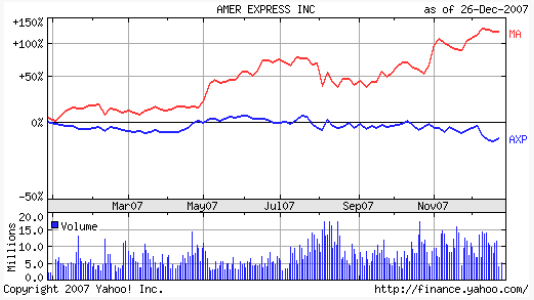
Medical Image Diagnoses

Image Processing

$$\begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n_{tr}1} & \cdots & x_{n_{tr}p} \end{pmatrix}$$



DNA Microarray Analysis



Stock market prediction

Why Learning?

- Understanding
- Prediction

Contents

- Fundamental framework; Statistical Decision Theory (SDT):
  - SDT; case of regression.
  - SDT; case of classification.
  - Statistical Learning from data, and two subfields.

- Design:
  - Different methods for regression in one picture.
  - Different methods for classification in one picture.

- Assessment:
  - Model Complexity and Bias-Variance tradeoff.
  - Cover's Theorem
  - Different measures.
  - Different paradigms.
  - Different estimators.

- Concluding Remarks

Different design methods in one picture.

In general,

$Y \in \mathbb{R}, X \in \mathbb{R}^p$; i.e., $X = (X_1, \ldots, X_p)$.

$Y \mid X = \mathrm{E}[\, Y \mid X \,] + \varepsilon$; and $\varepsilon \sim$ r.v.

1- Linear Models (LM)

Remember, the best regression function is

$$\eta^*(X) = \mathrm{E}[\, Y \mid X \,].$$

The assumption of the LM is that:

$$\eta^*(X) = E[Y \mid X] = \beta' X_{new}, \text{where}$$

$$X_{new} = (f_1(X), f_2(X), \ldots, f_d(X))',$$

Example: $X = (X_1, X_2)'$, and

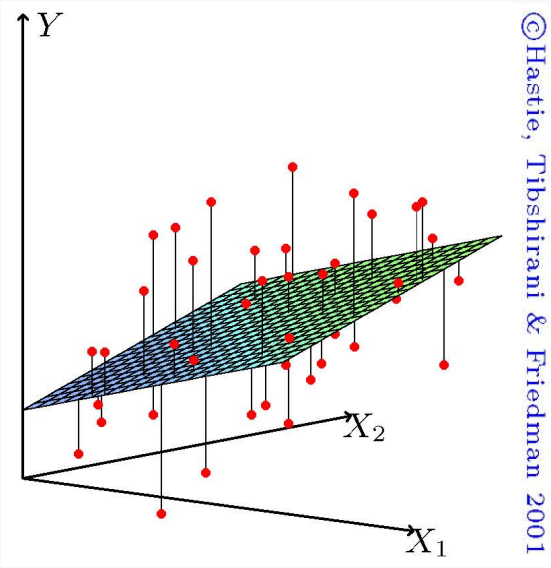
$$X_{new} = (X_1, X_2, X_1^2, X_1 X_2, \log X_2)'$$

LMS estimation of β:

$$\hat{\beta} = (\, \mathbf{X}'_{new}\, \mathbf{X}_{new}\,)^{-1} \mathbf{X}'_{new}\, \mathbf{y},$$

$$\mathbf{X} = \begin{pmatrix} \leftarrow\ x_1\ \rightarrow \\ \vdots \\ \leftarrow\ x_{n_{tr}}\ \rightarrow \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n_{tr}1} & \cdots & x_{n_{tr}p} \end{pmatrix},$$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_{n_{tr}} \end{pmatrix}, \text{and } \mathbf{X}_{new} = \begin{pmatrix} \leftarrow\ x_{1new}\ \rightarrow \\ \vdots \\ \leftarrow\ x_{n_{tr}new}\ \rightarrow \end{pmatrix}$$



Different design methods in one picture.

The LM assumption is true if

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim N(\mu, \Sigma), \text{ where } \mu = \begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

$$E[Y|X] = \mu_Y + \Sigma_{12}\Sigma_{22}^{-1}(X - \mu_X)$$

$$= \underbrace{(\mu_Y - \Sigma_{12}\Sigma_{22}^{-1}\mu_X)}_{\alpha} + \underbrace{\Sigma_{12}\Sigma_{22}^{-1}}_{\beta} X$$

Learning here is nothing but estimating μs and Σs;

After estimation, the linear model will be very close to $\eta^*(X)$;

yet not optimal unless $n_{tr} \to \infty$, only by then $\eta(X) \to \eta^*(X)$

Different design methods in one picture.

2- Generalized Linear Models (GLM)

More generally than in LM, we can assume

$g(E[\,Y \mid X\,]) = X'\beta$, for suitable link function g.

If we choose the logit function : $g(\mu) = \log \dfrac{\mu}{1-\mu}, \quad 0 < \mu < 1$, then

$\eta(X) = \dfrac{\exp(X'\beta)}{1+\exp(X'\beta)}$, which constrains that $0 < \eta(X) < 1$

3- Logistic Regression

$\Pr[\,\omega_1 \mid X = x\,] = \dfrac{\exp(\,x'\beta)}{1+\exp(\,x'\beta)}$, and $\Pr[\,\omega_2 \mid X = x\,] = \dfrac{1}{1+\exp(\,x'\beta)}$

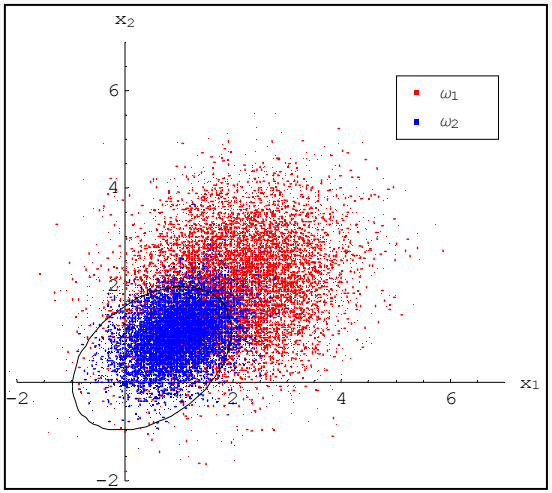
$l(\beta) = \log \prod_{i=1}^{n} \Pr[\,\omega_i \mid X = x_i\,]$ is the data likelihood

$= \sum_{i=1}^{n} \left\{ y_i x_i'\beta - \log(1+e^{x_i'\beta}) \right\}$ should be maximized for β;

the resulting equations are solved by iterative algorithm, e.g., Newton-Raphson

Different design methods in one picture.

4- Classification using the density functions



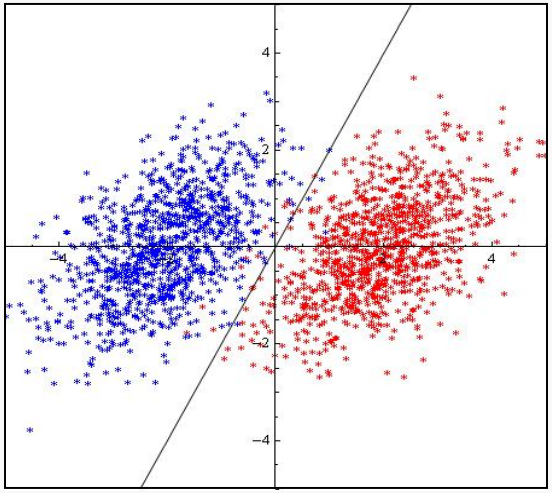
From Bayes' classifier, $\log \frac{f_X(X \mid \omega_1)}{f_X(X \mid \omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} th$

if we assume $f(X \mid \omega_k) \sim N(\mu_k, \Sigma_k)$, => QDA:

$$(x-\mu_1)'\Sigma_1^{-1}(x-\mu_1) - (x-\mu_2)'\Sigma_2^{-1}(x-\mu_2) \underset{\omega_2}{\overset{\omega_1}{\gtrless}} th$$

if we assume common cov. matrices Σ => LDA:

$$\underbrace{-\frac{1}{2}(\mu_1+\mu_2)'\Sigma^{-1}(\mu_1-\mu_2)}_{\alpha} + \underbrace{(\mu_1-\mu_2)'\Sigma^{-1}}_{\beta'} x \underset{\omega_2}{\overset{\omega_1}{\gtrless}} th$$



$\alpha + \beta' x \underset{\omega_2}{\overset{\omega_1}{\gtrless}} th$, which is the same form as Logistic Regression

But, more information here is available for α and β,

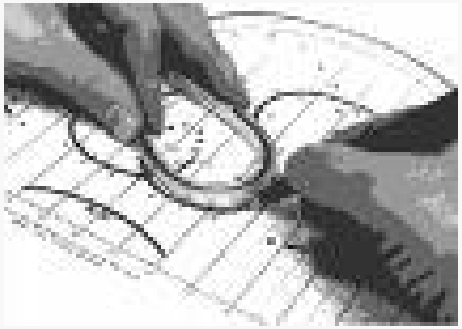
which comes from the Normality assumption => more accurate

"…It is our experience that the models give very similar results, even when LDA is used inappropriately, such as with qualitative predictors" [ESL]

Different design methods in one picture.

5- Nonparametric methods: No model is assumed.

- in contrast to parametric models.
- Let data speak; do not impose particular model.
- very useful when we cannot guess parametric form.
- **Smoothing** the response Y over the predictor X.



Different design methods in one picture.

5.1 K-Nearest Neighbor for regression:

$$\eta(x) = \sum_{i=1}^{n} y_i W_i(x),$$

$$W_i(x) = \begin{cases} \frac{1}{k} & i \in J_x = \{\, i : x_i \in N_k(x)\} \\ 0 & \text{otherwise} \end{cases}$$

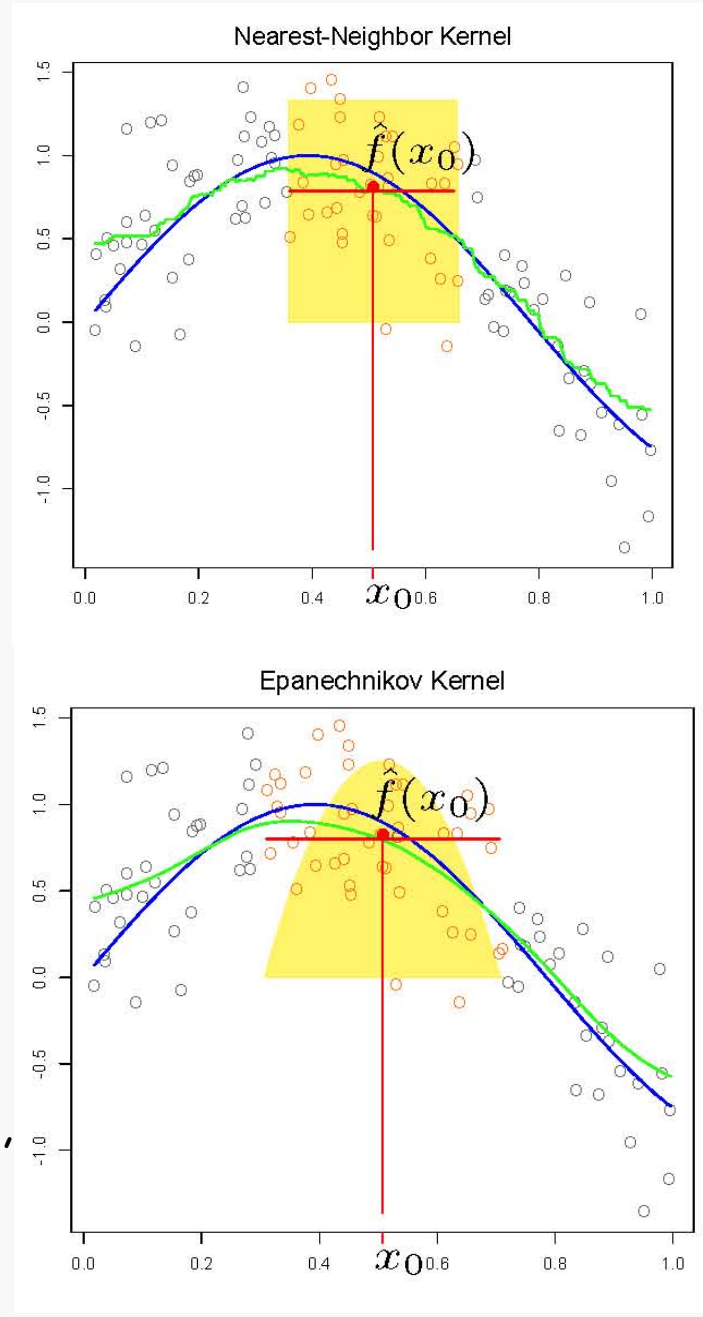
approximating $E(Y \mid X)$ by local averaging.

5.2 Kernel Smoothing for regression

$$\eta(x) = \sum_{i=1}^{n} y_i \left(\frac{K\left(\frac{x - x_i}{h_x}\right)}{\sum_{i'=1}^{n} K\left(\frac{x - x_{i'}}{h_x}\right)} \right)$$

approximating $E(Y \mid X)$ by weighted local averaging,

the smaller window size h_x the more complex $\eta(X)$



Different design methods in one picture.

5.3 K-Nearest Neighbor for Classification

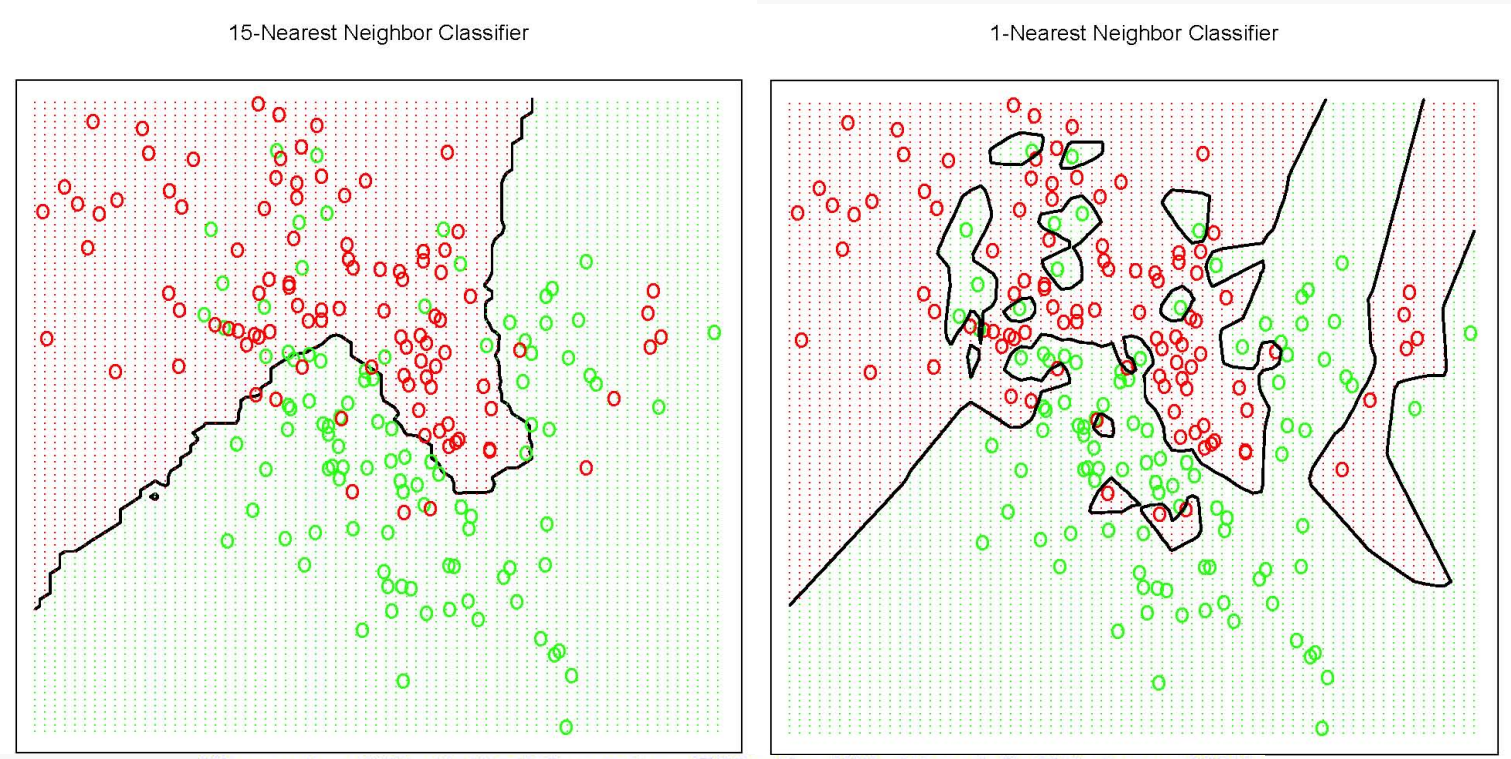
$$\Pr[\omega_j \mid x] = \frac{1}{n}\sum_{i=1}^{n} W_i(x) I_{\omega_i=\omega_j},$$

$$I_{cond} = \begin{cases} 1 & cond\ is\ True \\ 0 & cond\ is\ False \end{cases}$$

Approximating $\frac{\Pr[\omega_j \mid x]}{\Pr[\omega_i \mid x]}$ by local voting

Complexity decreases with K.

With $K = 1$ classifiers adapt too much to the training set.



Elements of Statistical Learning ©Hastie, Tibshirani & Friedman 2001

Different design methods in one picture.

5.4 Additive Models (AM)

$\eta(x) = \alpha + \sum_{i=1}^{p} f_i(X_i)$, where

Similar to LM: $\eta(x) = \alpha + \beta' X_{new} = \alpha + \sum_i \beta_i f_i(X)$ with two differences:

- Every term f_i is a function in only one dimension X_i
- Every term f_i has no particular parametric form.

Algorithm of estimating f_j using nonparametric smoothing (backfitting):

$\hat{\alpha} = \frac{1}{n}\sum_i y_i, f_j = 0 \,\forall\, i, j;$

do

$for(j = 1;\ j \le p;\ j++)$

$\hat{f}_j \leftarrow \mathrm{S}\Big[y - \hat{\alpha} - \sum_{k \neq j} \hat{f}_k \Big]$

$\hat{f}_j \leftarrow \hat{f}_j - Average\Big[\hat{f}_j \Big]$

end

$while\{\hat{f}_{j+1} - \hat{f}_j > \varepsilon\}$

Different design methods in one picture.

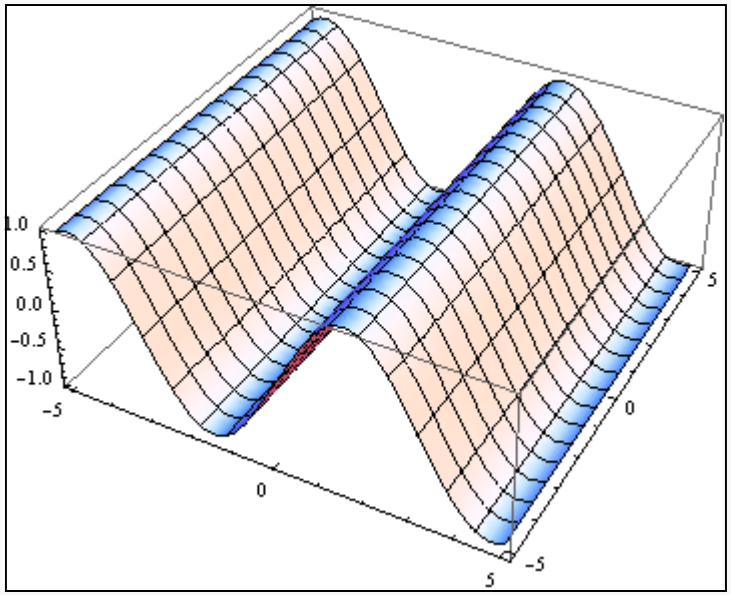
5.5 Generalized Additive Models (GAM)

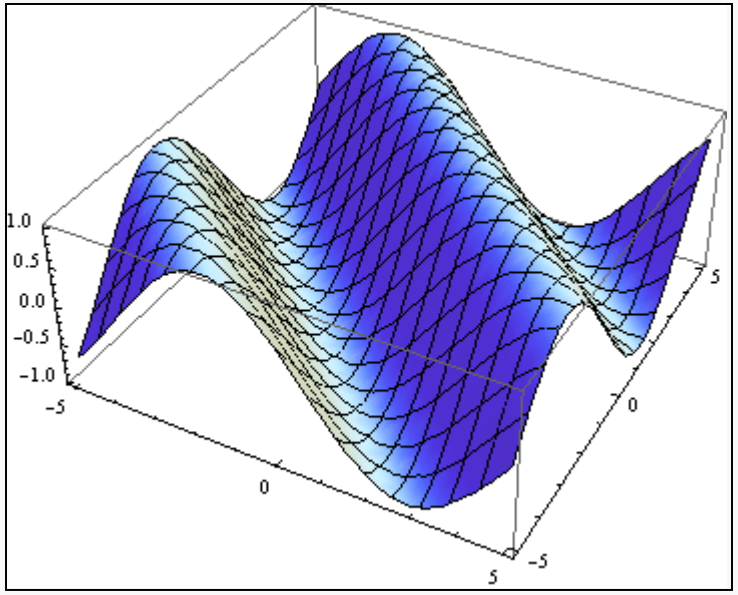
$$g(\eta(x)) = \alpha + \sum_{i=1}^{p} f_i(X_i)$$

GAM to AM is exactly as GLM to LM, i.e., modeling $g(\eta(x))$ instead of $\eta(x)$

Anologously to GLM:

we can set $g(\mu) = \log \frac{\mu}{1-\mu}$ and have a nonparametric Logistic regression





The function $\sin(X_1)$ is rotated by $\frac{\pi}{4}$ to prodcue $\sin(\frac{X_1}{\sqrt{2}} + \frac{X_2}{\sqrt{2}})$, Which is not additive.

Different design methods in one picture.

5.6 Projection Pursuit Regression (PPR), a remedy for the direction problem

$$\eta(x) = \sum_{i=1}^{k} f_i(\alpha_i' x)$$

g_i is a general smoothing technique, and α_i is the best projecting direction.

by setting $\alpha_1 = (1,0,...), \alpha_2 = (0,1,0,...)$; and so on it reduces to AM, i.e.,

$$\eta(x) = \sum_{i=1}^{k} f_i(x_i)$$

For the case of classification, a link function can be introduced to $\eta(X)$, i.e.,

$$g(\eta(x)) = \sum_{i=1}^{k} f_i(\alpha_i' x), \text{where } g(\mu) = \log \frac{\mu}{1-\mu}$$

This makes the projection pursuits the most general GAM.

Different design methods in one picture.

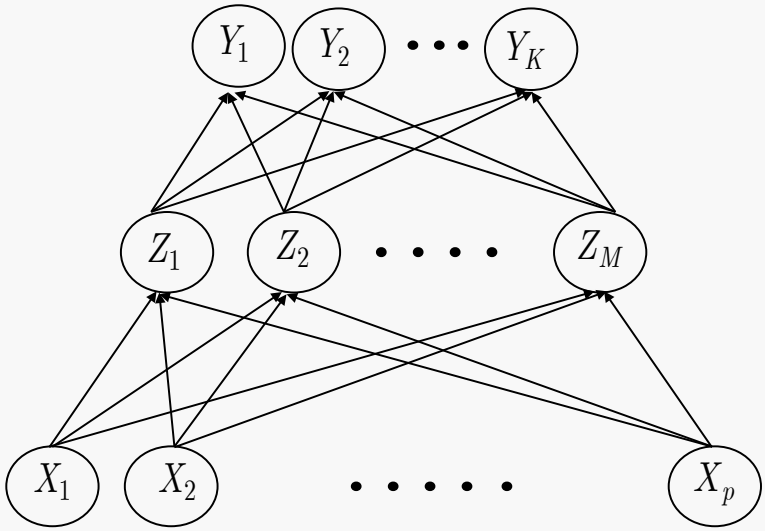
5.7 Neural Networks and the nice connection

$$Z_m = \sigma(\alpha_{om} + \alpha_m' X), \quad m = 1,2,\ldots,M,$$

$$Y_k = g_k\left(\beta_{0k} + \sum_{m=1}^{M} \beta_{mk} Z_m\right), \quad k = 1,2,\ldots,K,$$

$$\sigma(\mu) = \frac{1}{1+e^{-\mu}}$$

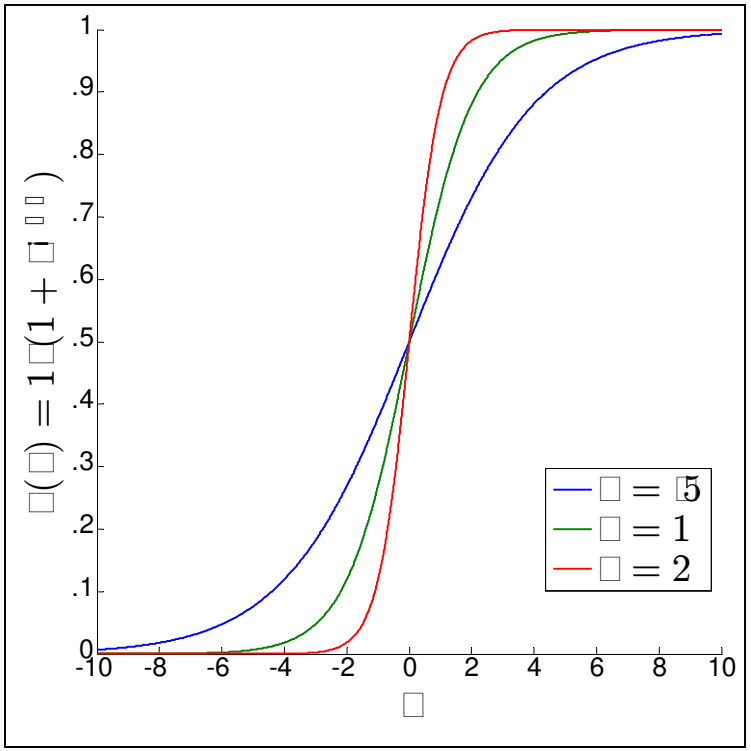
$$Y_k = T_k\left(\sum_{m=1}^{M} \sigma_{mk}(\alpha_m' X)\right), \text{ which is PPR!}$$



More over, PPR is more general, since the functions f_i are freely selected by smoothing techniques rather than imposing them to be the sigmoid function σ.

$T_k(\mu) = \mu$ => regression, and Y_k is the response

$T_k(\mu) = \dfrac{e^{\mu_k}}{\sum_k e^{\mu_k}}$ => classification, and $Y_k = \Pr(\omega_k \mid X)$



Avoid the fallacy; cont.

Some practitioners think of some decision rules (e.g., neural networks) as superior to all other rules.

The only one with this superiority is the theoretical Bayes' rule, which is unique. Any other method performs well ONLY IF the assumption made for it is satisfied in the dataset of interest.

" There has been a great deal of hype surrounding neural networks, making them seem magical and mysterious. As we make clear in this section, they are just nonlinear statistical models, much like the projection pursuit regression model discussed above." [ESL]

[ESL] *Hastie, T., R. Tibshirani, and J.H. Friedman (2001), The elements of statistical learning: data mining, inference, and prediction. Springer series in statistics., New York: Springer*

Different design methods in one picture.

Some other methods:

Basis expansion
Splines
Classification and Regression Tress (CART)
Boosting and Additive Trees
Support Vector Machine (SVM)
Density function estimation

Crucial FAQ:

1. Which method to apply on my dataset?

2. Even I chose a method, what about complexity?

in LM, if $X = (X_1, X_2)$, should I involve ${X_1}^2, {X_2}^2, {X_1}^3, ...$

in Kernel smoothing, should I choose a very narrow window size h?

in KNN, what is appropriate K?

in NN, how many layers?

Contents

- Fundamental framework; Statistical Decision Theory (SDT):
  - SDT; case of regression.
  - SDT; case of classification.
  - Statistical Learning from data, and two subfields.

- Design:
  - Different methods for regression in one picture.
  - Different methods for classification in one picture.

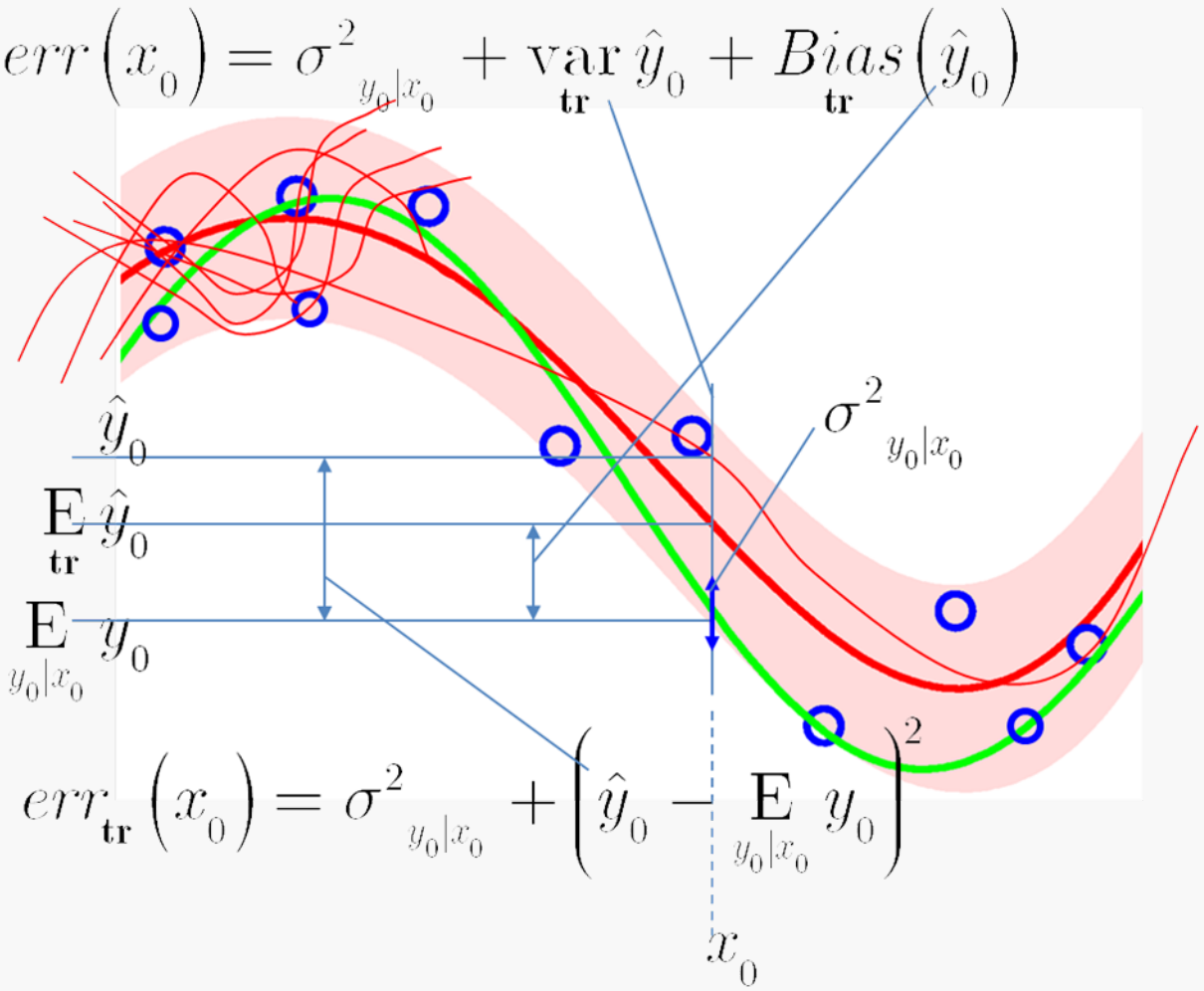
- Assessment:
  - Model Complexity and Bias-Variance tradeoff.
  - Cover's Theorem
  - Different measures.
  - Different paradigms.
  - Different estimators.

- Concluding Remarks

Model Complexity and Bias-Variance tradeoff

$$R(\eta_{\mathbf{tr}}) = \mathrm{E}\left[\eta_{\mathbf{tr}}(x_0) - y_0\right]^2$$

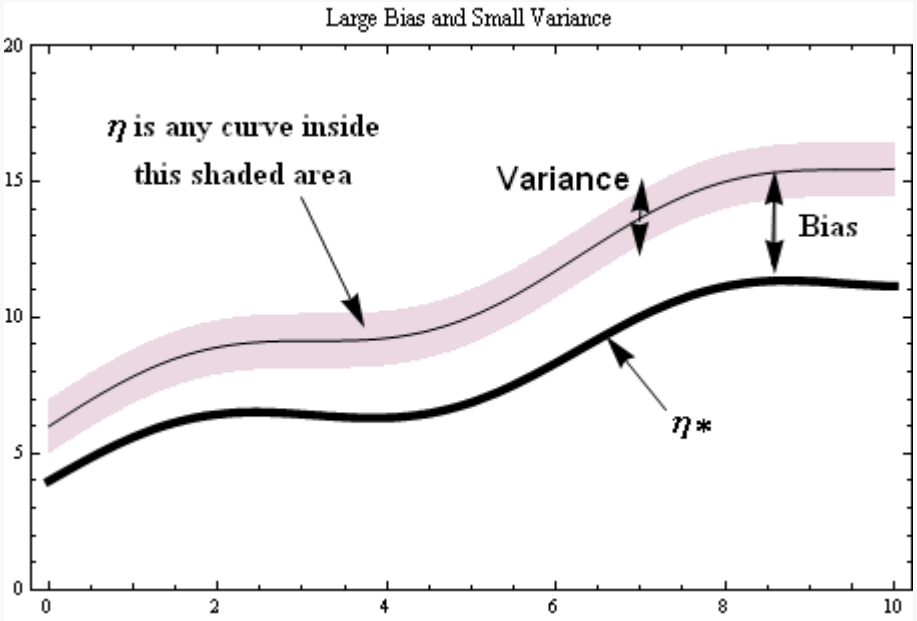
$$= \underbrace{\mathrm{E}\left(\eta^*(x_0) - y_0\right)^2}_{Bayes\ Risk = \mathrm{E}\,\mathrm{var}[Y|X]} + \underbrace{\left\{\mathrm{E}\,\eta_{\mathbf{tr}}(x_0) - \eta^*(x_0)\right\}^2}_{Bias^2} + \underbrace{\mathrm{E}[\eta_{\mathbf{tr}}(x_0) - \mathrm{E}\,\eta_{\mathbf{tr}}(x_0)]^2}_{Variance}$$

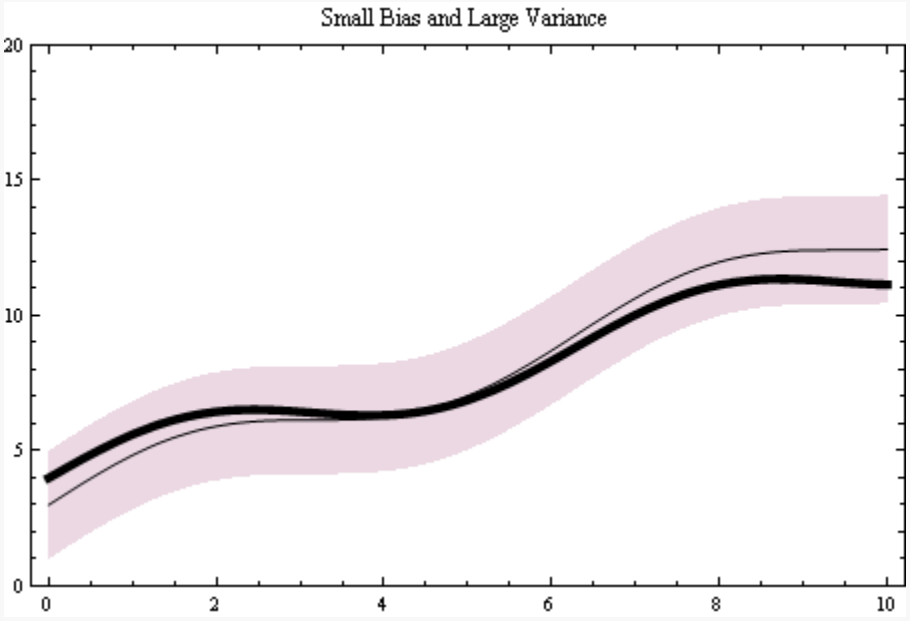


Model Complexity and Bias-Variance tradeoff

$$R(\eta_{\mathbf{tr}}) = \mathrm{E}\left[\eta_{\mathbf{tr}}(x_0) - y_0\right]^2$$

$$= \underbrace{\mathrm{E}\left(\eta^*(x_0) - y_0\right)^2}_{Bayes\ Risk = \mathrm{E}\,\mathrm{var}[Y|X]} + \underbrace{\left\{\mathrm{E}\,\eta_{\mathbf{tr}}(x_0) - \eta^*(x_0)\right\}^2}_{Bias^2} + \underbrace{\mathrm{E}[\eta_{\mathbf{tr}}(x_0) - \mathrm{E}\,\eta_{\mathbf{tr}}(x_0)]^2}_{Variance}$$





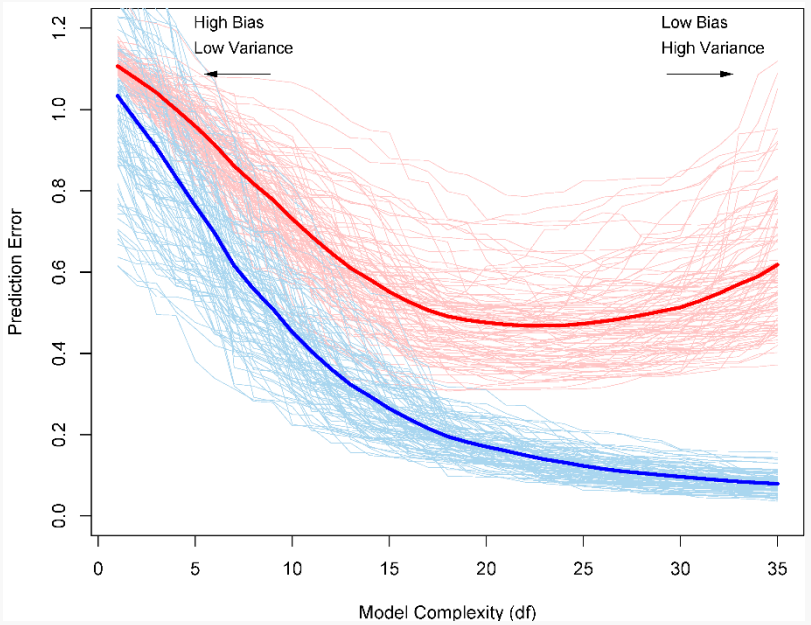
Model Complexity and Bias-Variance tradeoff

Remember: the performance is measured on the unseen data NOT the training data.

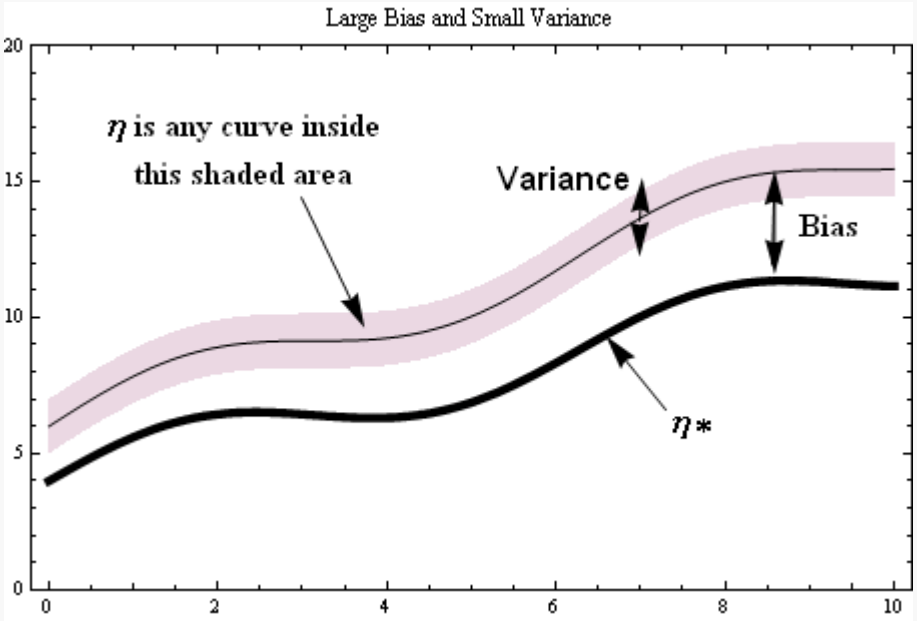
Bias decreases with complexity, because you adapt too much to the data whose average is the best function.

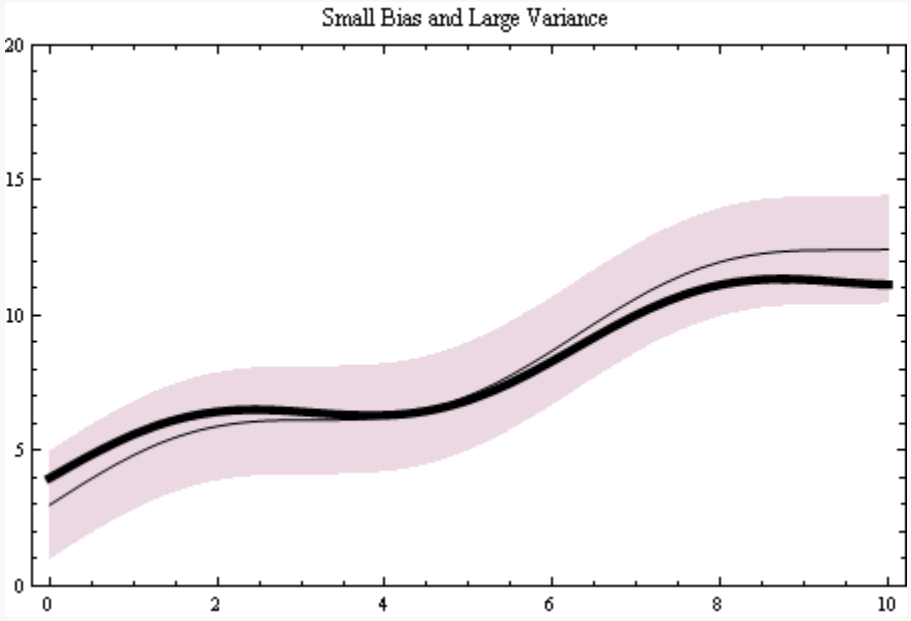
Variance increases with complexity, because you adapt too much to the data which is variable

There is an optimal complexity!

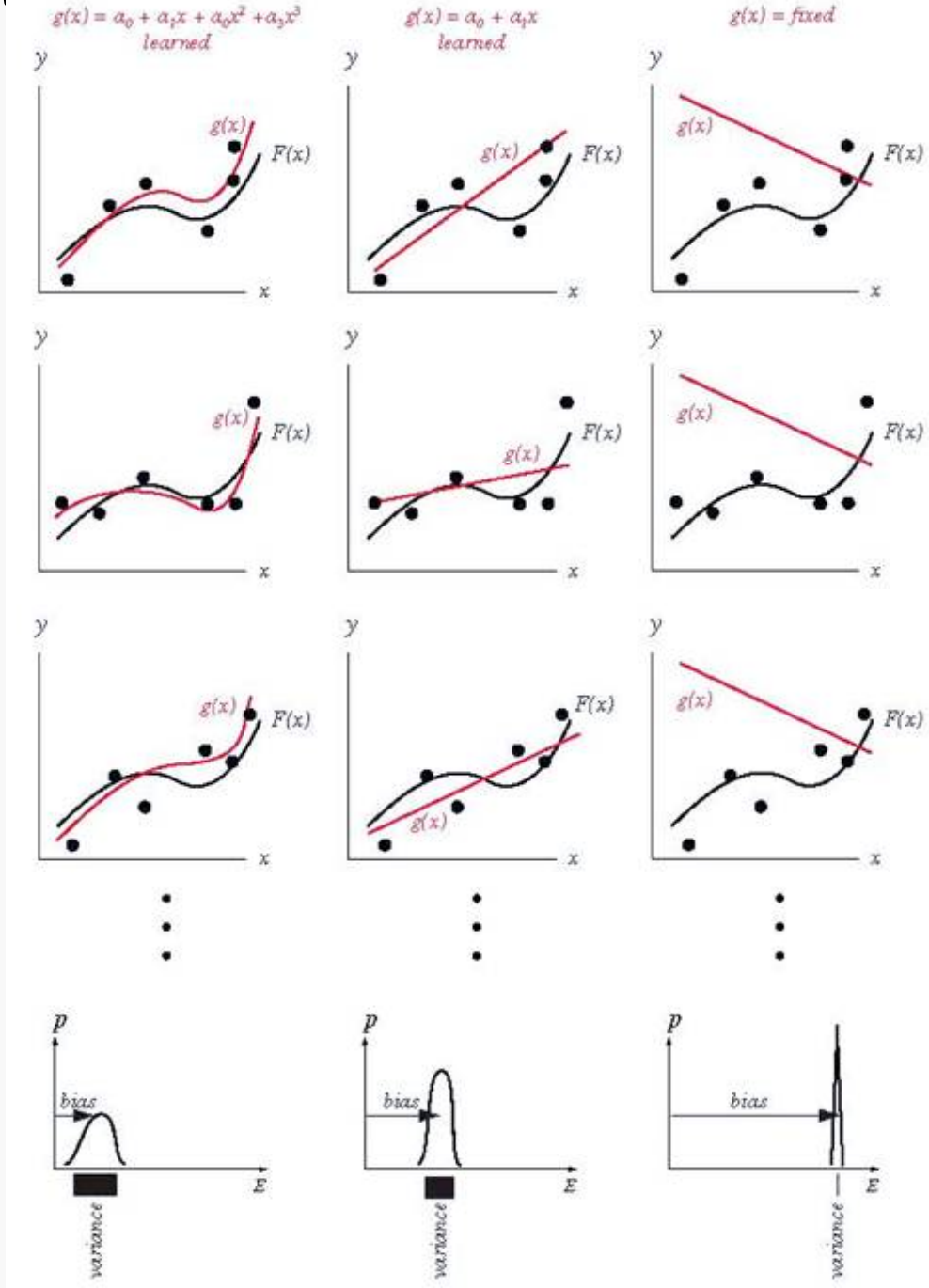


Elements of Statistical Learning ©Hastie, Tibshirani & Friedman 2001

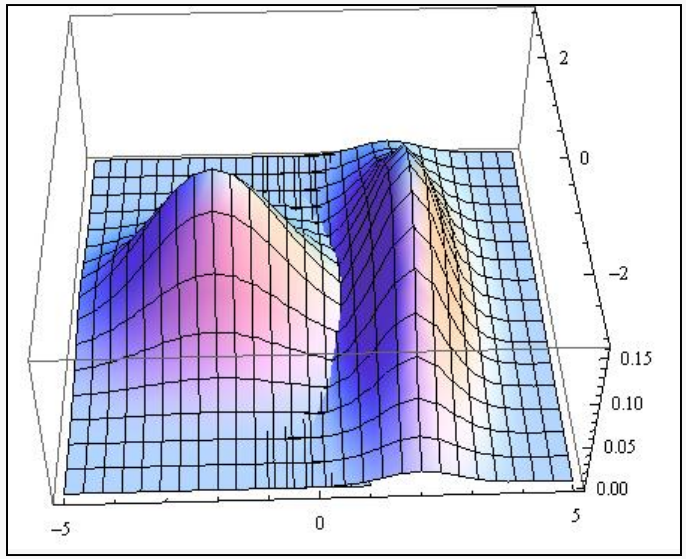


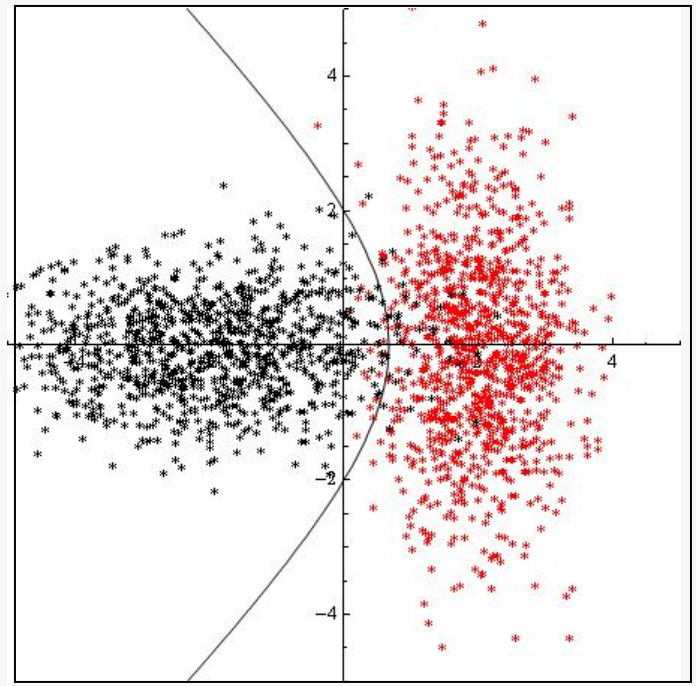


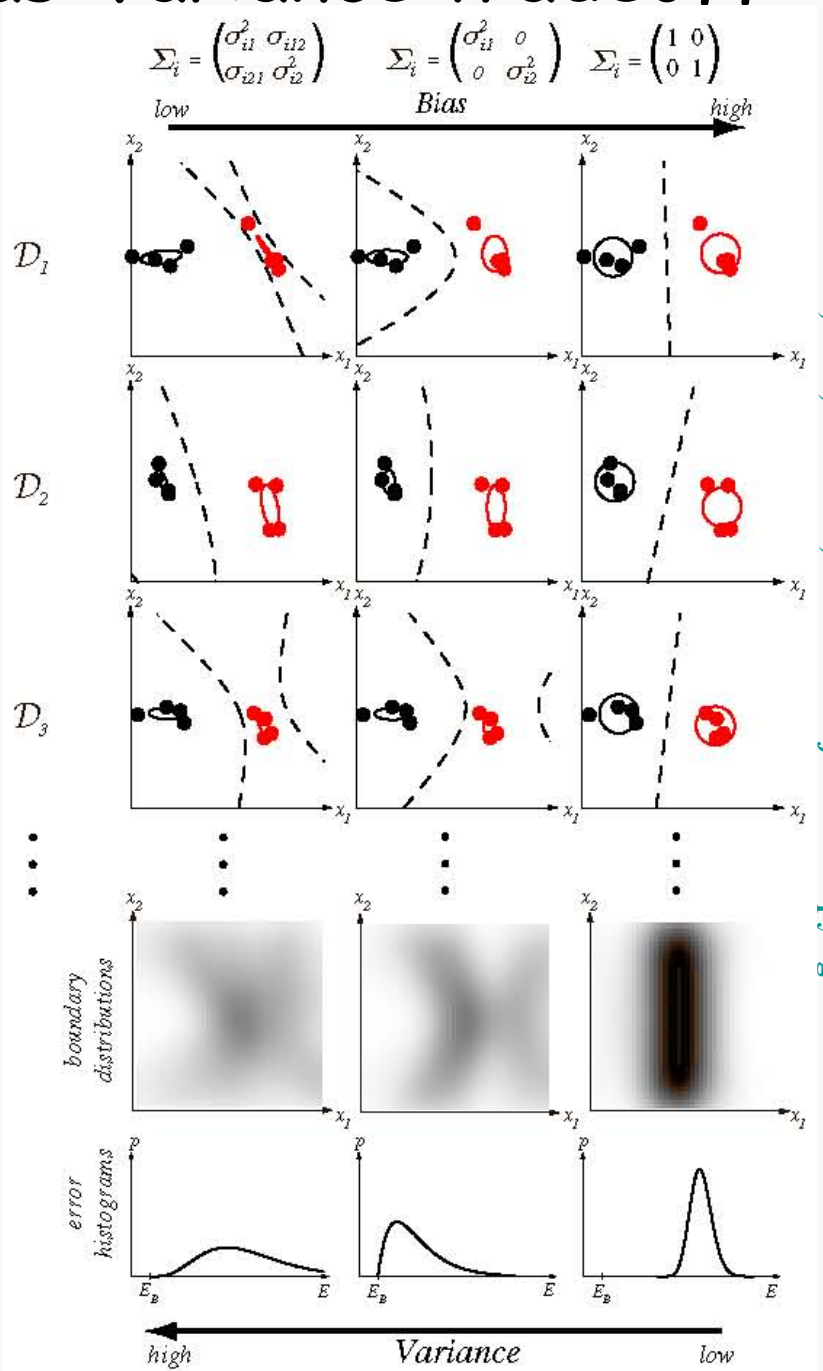
Model Complexity and Bias-Variance tradeoff



Model Complexity and Bias-Variance tradeoff

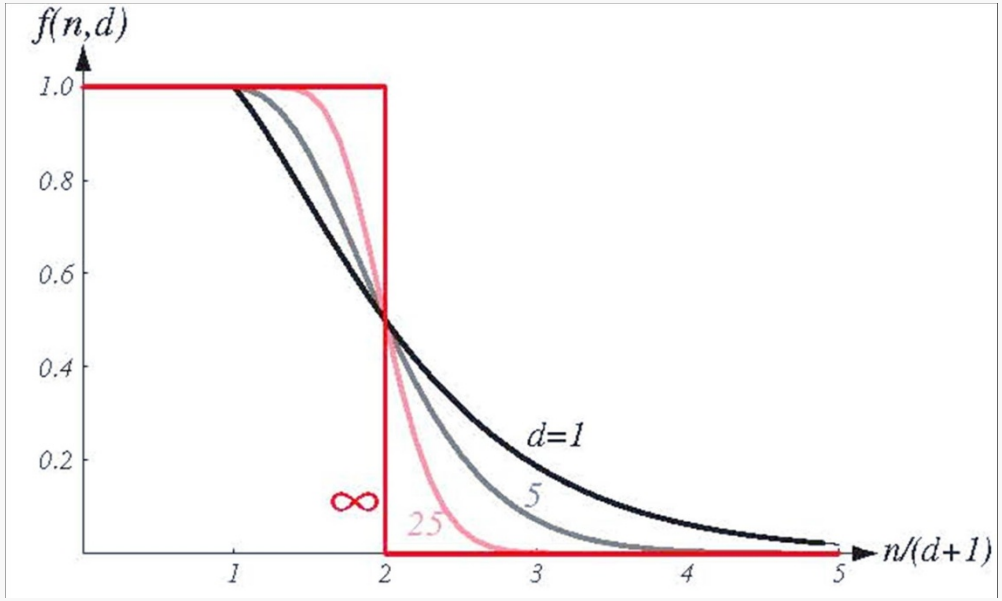


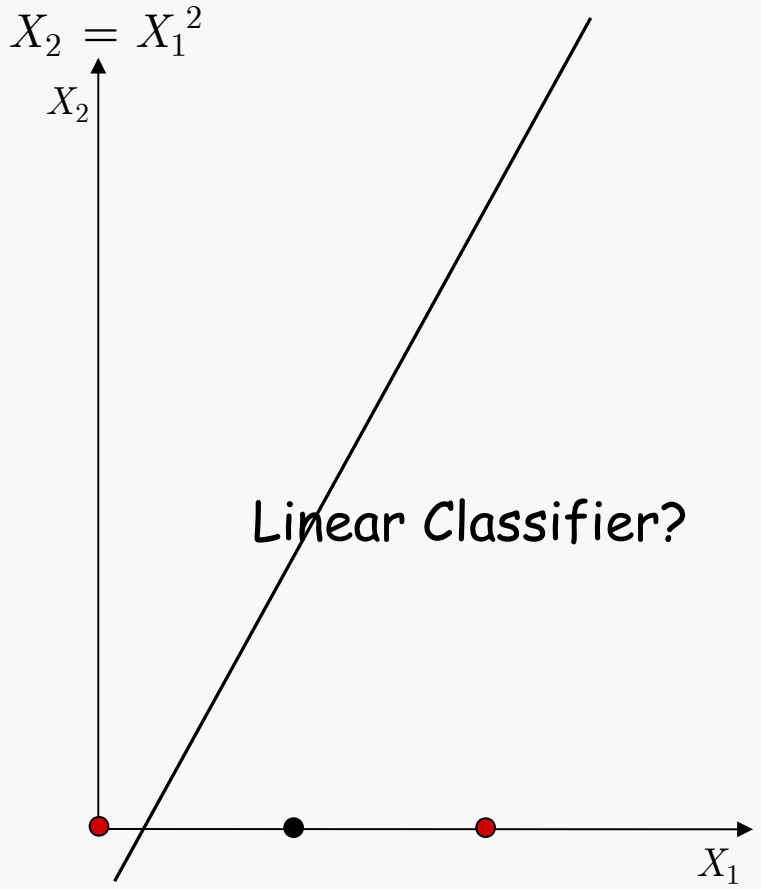




Connection to Cover's theorem:

Cover, T. M. (1965). "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition." IEEE Transactions on Electronic Computers.

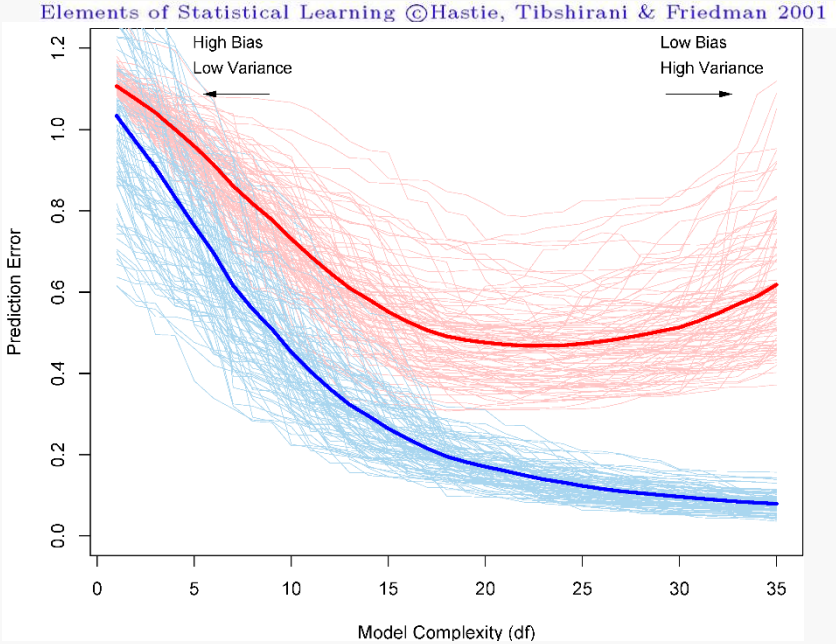


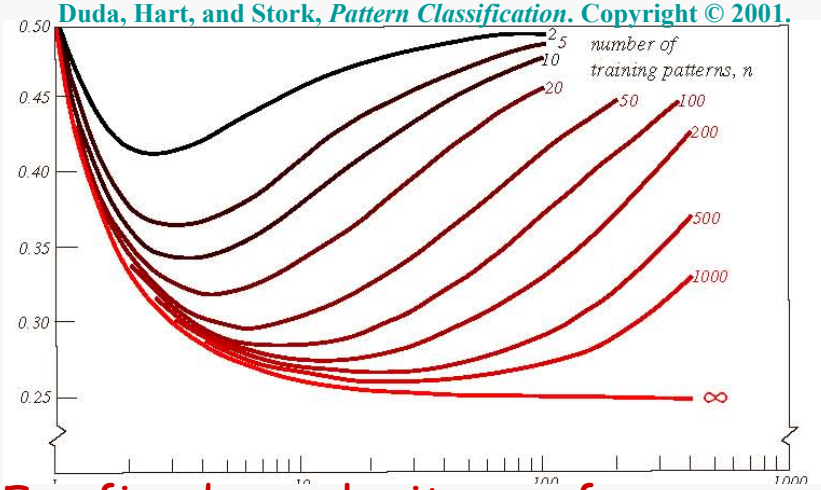


For the same number of observations, the linear classification is easier at higher dimensions =>
less apparent error NOT true error; so in fact you are not learning. **This is what is called ill-posed problems.**

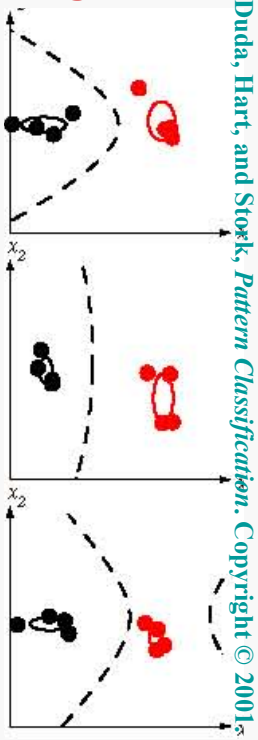
At higher dimensions the problem is geometrically easier (the apparent data) not probabilistically separable (the unseen data).

Reinforcement, and introduction to assessment





For fixed complexity, performance almost improves with training set size

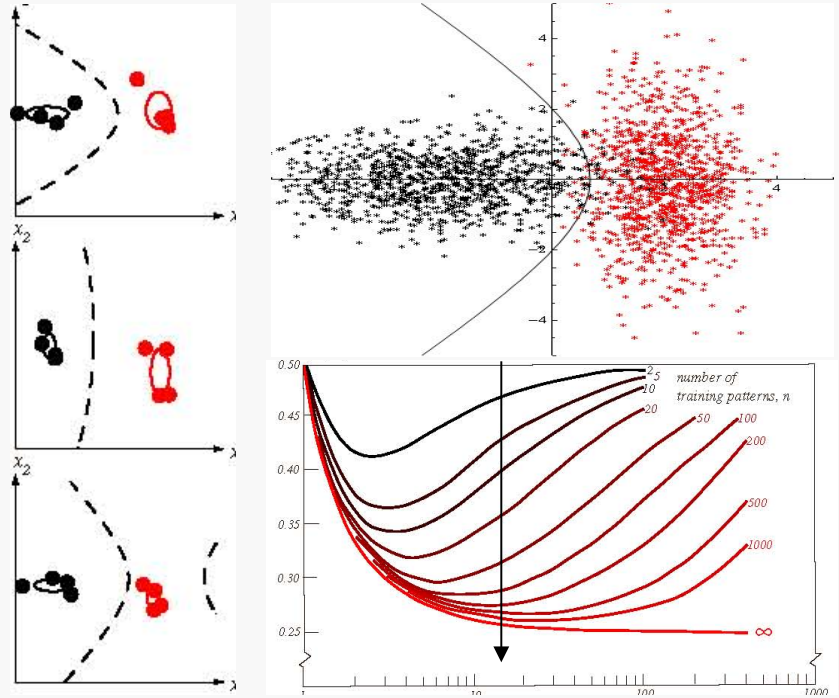


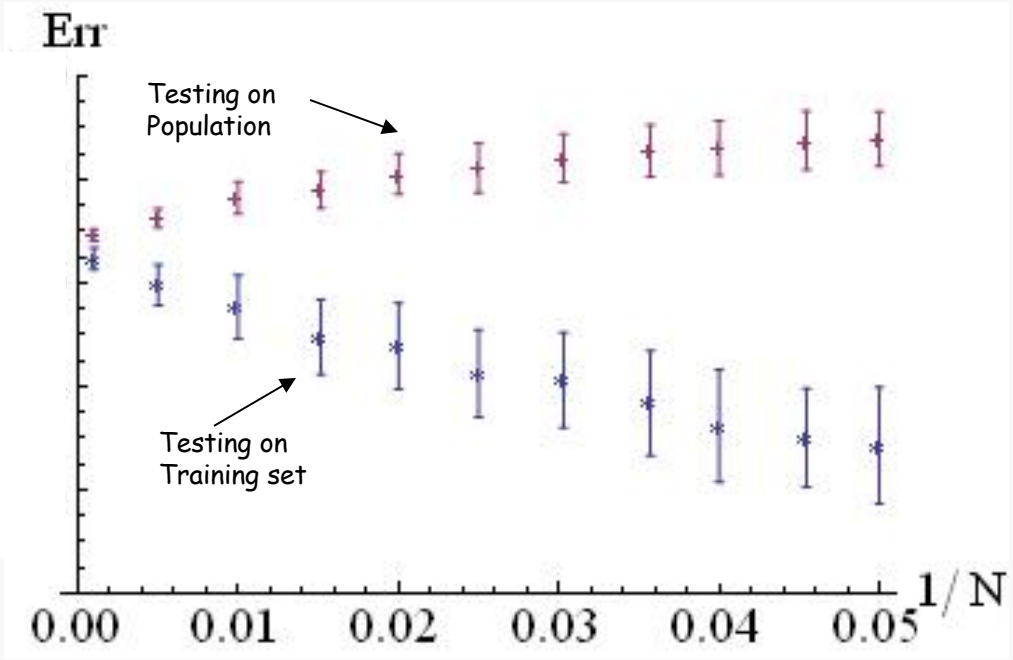
- Some suffice with estimating the performance of $\eta_{\mathbf{tr}}$ (conditional performance), e.g., $Err_{\mathbf{tr}}$
- We have to estimate the mean performance, i.e., $\mathrm{E}_{\mathbf{tr}} Err_{\mathbf{tr}}$
- We have to estimate the uncertainty in any estimation
- What is performance?

Performance vs. Training-set Size

The true $Err_{\mathbf{tr}} \sim R.V.$(conditional on $\mathbf{tr}$) $\Rightarrow$ $\mathrm{E}(Err_{\mathbf{tr}})$ & $VAR(Err_{\mathbf{tr}})$

$\mathbf{ts} = \{t_1, \ldots, t_{n_{ts}}\}$, where $n_{ts} \to \infty \equiv$ known $f_{X|\omega_1}$ and $f_{X|\omega_2}$





- $\mathbf{ts} = \mathbf{tr}$ (apparent performance or "resubstitution")
- $\mathbf{ts}$ and $\mathbf{tr}$ are iteratively produced by bootstrapping

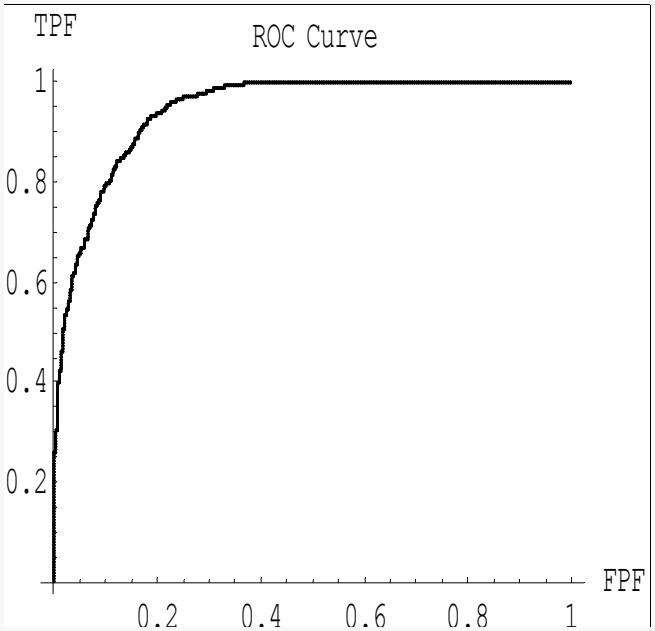
  Or Cross-Validation (Paradigm 1)
- $\mathbf{ts} = \{t_1, \ldots, t_{n_{ts}}\}$ separate from $\mathbf{tr}$, and n_{ts} is finite (Paradigm 2)

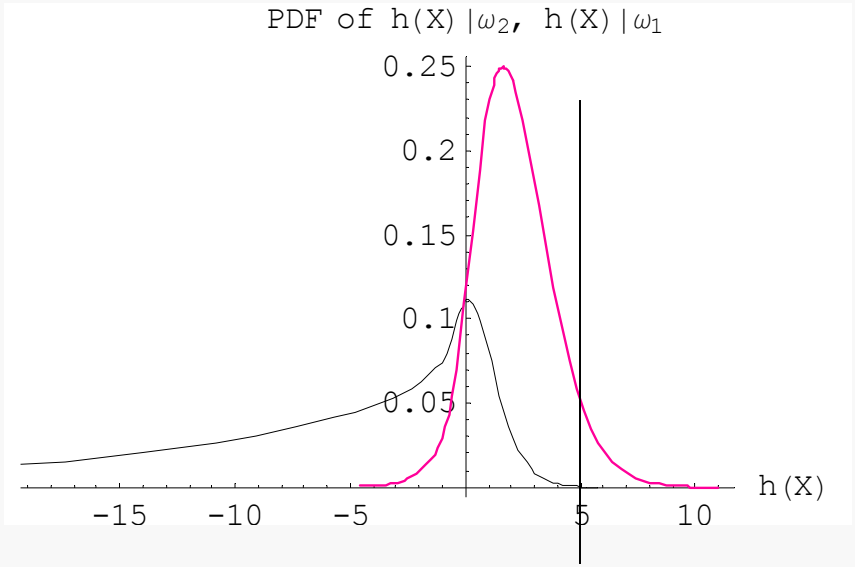
$\Sigma_1 = \Sigma_2 = \mathbf{I}$, $\mu_1 = \mathbf{0}$

$\mu_2 = 0.34\mathbf{1}$, $d = 7$

$Mahalanobis\, dist.^2 = 0.8$

Performance Measures for Classifiers: ROC





$$h(X) \underset{\omega_2}{\overset{\omega_1}{\gtrless}} th, \text{where } h(X) = \log \frac{f_X(X = x \mid \omega_1)}{f_X(X = x \mid \omega_2)}, \text{and } th = \log \frac{\Pr(\omega_2)(\, c_{21})}{\Pr(\omega_1)(\, c_{12})}$$

- True AUC:

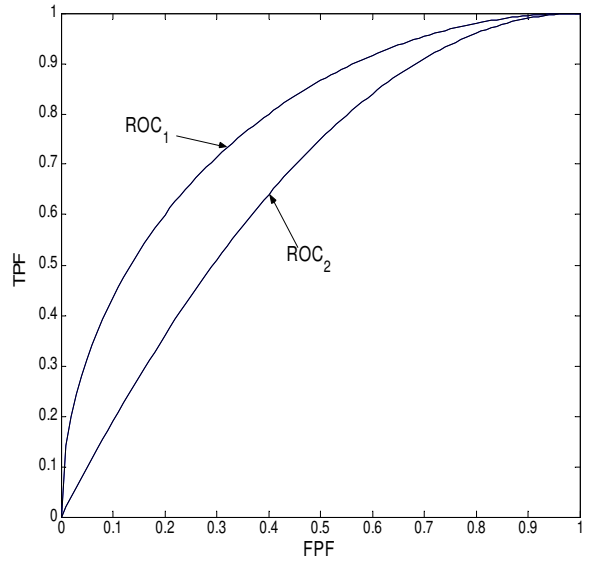
$$AUC = \int_0^1 TPF \; d(FPF)$$

- it is easy to show that:

$$AUC = \Pr\{\, h(X \mid \omega_1) > h(X \mid \omega_2)\}$$

High AUC means high performance.

High Err means low performance.

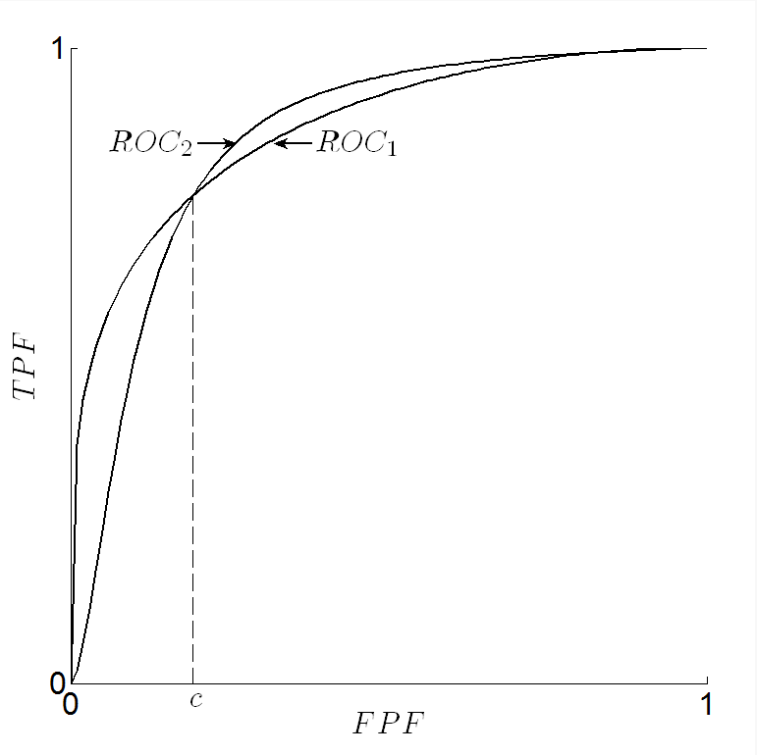


Partial Area under the ROC Curve (PAUC): not mature yet; still under research;

Parametric Definition:

$$PAUC_{\mathbf{tr}}(th_c) = \int_{u=\infty}^{u=th_c} TPF_{\mathbf{tr}}(u)\, dFPF_{\mathbf{tr}}(u)$$

Call th_c the cutoff threshold



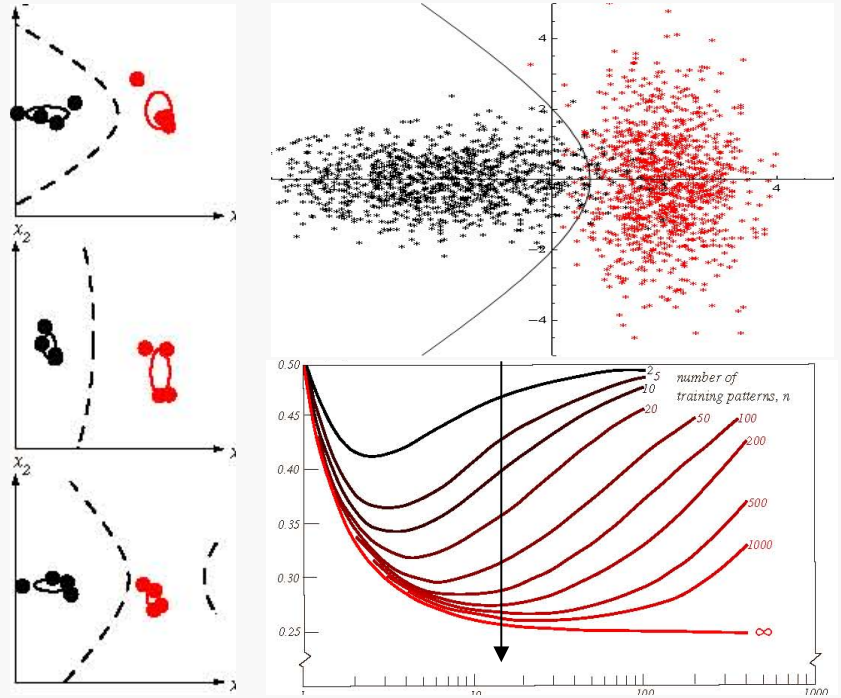
it can be proven that:

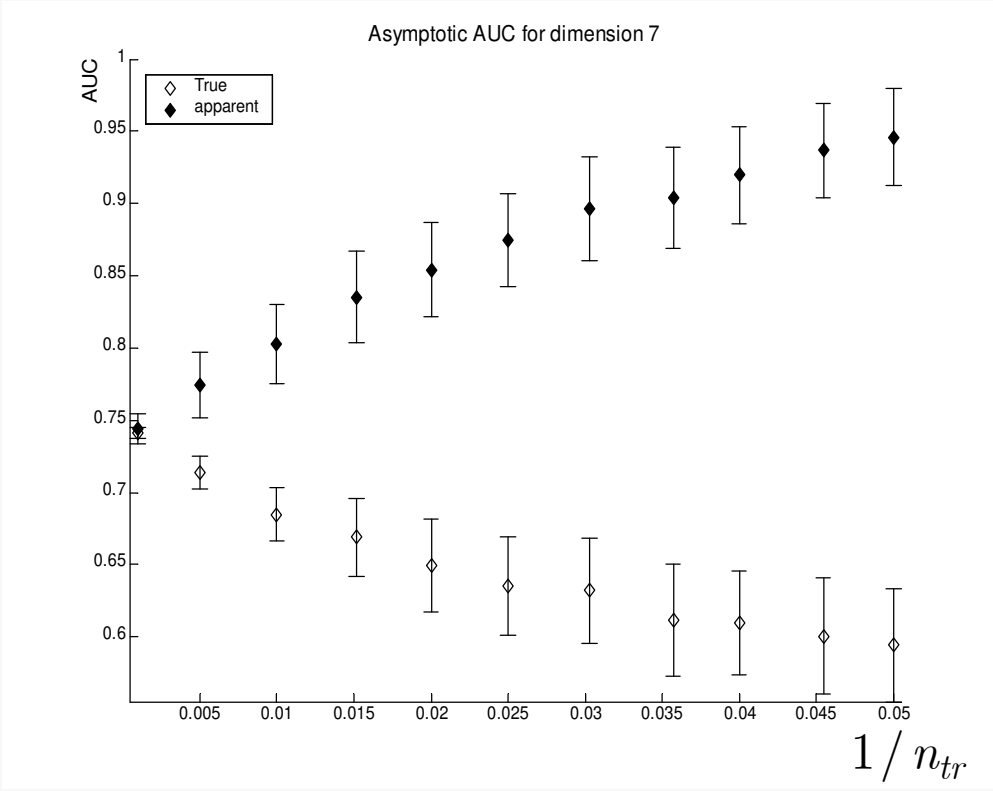
$$PAUC_{\mathbf{tr}}(th_c) = \Pr\{\, h_{\mathbf{tr}}(x \mid \omega_1) > h_{\mathbf{tr}}(x \mid \omega_2) > th_c \,\}$$

Performance vs. Training-set Size

The true $AUC_{\mathbf{tr}} \sim RV.(\text{conditional on } \mathbf{tr}) \quad \Rightarrow \quad \mathrm{E}(AUC_{\mathbf{tr}})\ \&\ VAR(AUC_{\mathbf{tr}})$

$\mathbf{ts} = \{t_1, \ldots, t_{n_{ts}}\}$, where $n_{ts} \to \infty \equiv$ known $f_{X|\omega_1}$ and $f_{X|\omega_2}$





- $\mathbf{ts} = \mathbf{tr}$ (apparent performance or "resubstitution")
- $\mathbf{ts}$ and $\mathbf{tr}$ are iteratively produced by bootstrapping
  Or Cross-Validation (Paradigm 1)
- $\mathbf{ts} = \{t_1, \ldots, t_{n_{ts}}\}$ separate from $\mathbf{tr}$, and n_{ts} is finite (Paradigm 2)

$\Sigma_1 = \Sigma_2 = \mathbf{I},\ \mu_1 = \mathbf{0}$
$\mu_2 = 0.34\mathbf{1},\ d = 7$
$Mahalanobis\, dist.^2 = 0.8$

K-fold Cross Validation (KCV): for model selection and assessment

$$\widehat{err}_{\mathbf{tr}}^{(KCV)}(\hat{f},\lambda) = \frac{1}{N}\sum_{i} L\left(y_i, \hat{f}^{-\mathcal{K}(i)}(x_i,\lambda)\right)$$

```
KCV ALGORITHM for model M

    Divide the N-observation dataset to...
        K partitions, each has N/K;

    For k = 1 : K
        train on all data except partition k;
        test on partition k;
        save the N/K predictions;
    end

    collect the K * N/K predictions;

    Estimate your error;
```

Stone, M., (1974). "Cross-Validatory Choice and Assessment of Statistical Predictions". *Journal of the Royal Statistical Society. Series B (Methodological)*

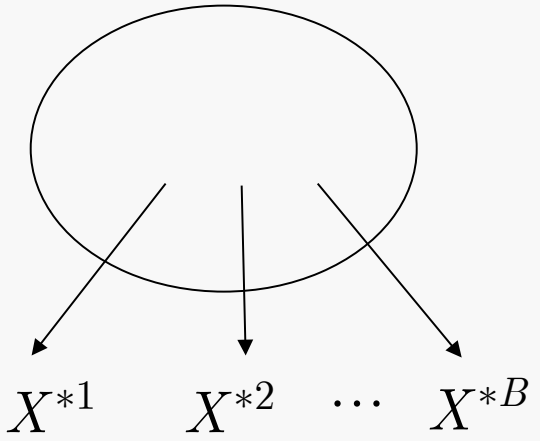
```
KCV ALGORITHM for model selection

    for m = 1 : M
        Err[m] = error of model m...
            estimated using KCV;
    end

    find the minimum Err and its model;
```

Introduction to the Bootstrap

Without Bootstrap: Original Sample

Unknown Distribution | Observed Sample | Statistic of Interest

$$F \rightarrow \quad \mathbf{x} = (x_1, x_2, \ldots x_n) \rightarrow \quad \hat{\theta} = s(\mathbf{x})$$



With Bootstrap: Many Replicated Samples

Unknown Distribution | Observed Sample | Bootstraps Estimates

$\hat{F}$

$$\mathbf{x}^{*1} = \left(x_1^{*1}, x_2^{*1}, \ldots x_n^{*1} \right) \rightarrow \hat{\theta}^{*1} = s\left(\mathbf{x}^{*1} \right)$$

$$\mathbf{x}^{*2} = \left(x_1^{*2}, x_2^{*2}, \ldots x_n^{*2} \right) \rightarrow \hat{\theta}^{*2} = s\left(\mathbf{x}^{*2} \right)$$

$$\vdots$$

$$\mathbf{x}^{*b} = \left(x_1^{*b}, x_2^{*b}, \ldots x_n^{*b} \right) \rightarrow \hat{\theta}^{*b} = s\left(\mathbf{x}^{*b} \right)$$

$$\vdots$$

$$\mathbf{x}^{*B} = \left(x_1^{*B}, x_2^{*B}, \ldots x_n^{*B} \right) \rightarrow \hat{\theta}^{*B} = s\left(\mathbf{x}^{*B} \right)$$

Efron, B. & Tibshirani, R., (1997). "Improvements on Cross-Validation: The .632+ Bootstrap Method". *Journal of the American Statistical Association,* 92(438), pp.548–560.

Efron, B., (1983). "Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation". *Journal of the American Statistical Association,* 78(382), pp.316–331.

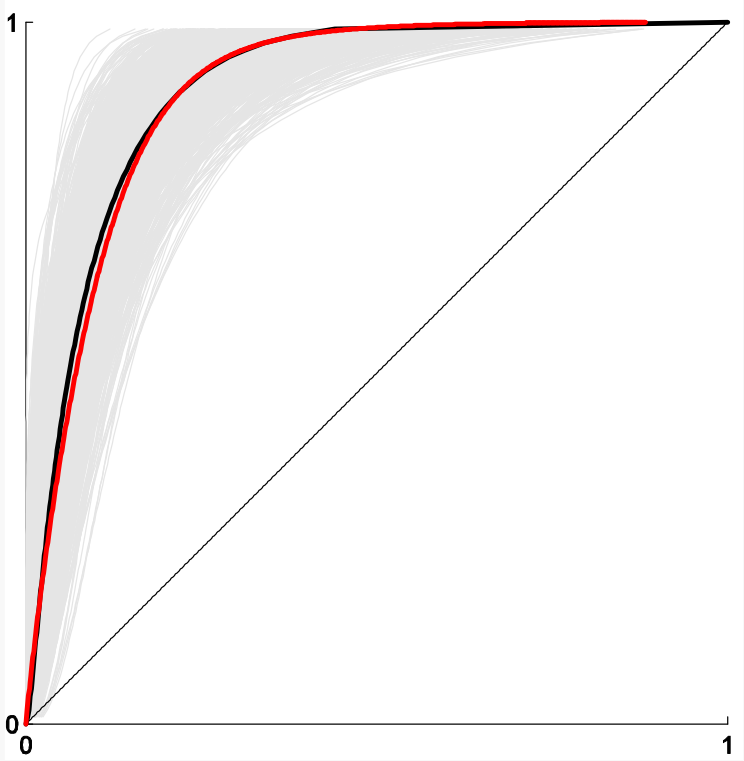
Parametric Assessment and whole ROC estimation

-We started by Err (one point on the curve)
- Then Nonparametric AUC estimation.

Now, how to estimate the whole curve

Metz's SW:
ROCKIT
PROPROC

Based on parametric fit



Metz, C.E., (1986). Statistical Analysis of ROC Data in Evaluating Diagnostic Performance. In *Multiple regression analysis: Applications in the health sciences*. pp. 365–384.

Metz, C. E., B. A. Herman, et al. (1998). "Maximum Likelihood Estimation of Receiver Operating Characteristic (ROC) Curves from Continuously-Distributed Data." Statistics In Medicine. 17(9), pp.1033–1053.

Metz, C. E. and X. Pan (1999). "Proper Binormal ROC Curves: Theory and Maximum-Likelihood Estimation." Journal of Mathematical Psychology. 4(2), pp.138–149.

Contents

- Fundamental framework; Statistical Decision Theory (SDT):
  - SDT; case of regression.
  - SDT; case of classification.
  - Statistical Learning from data, and two subfields.

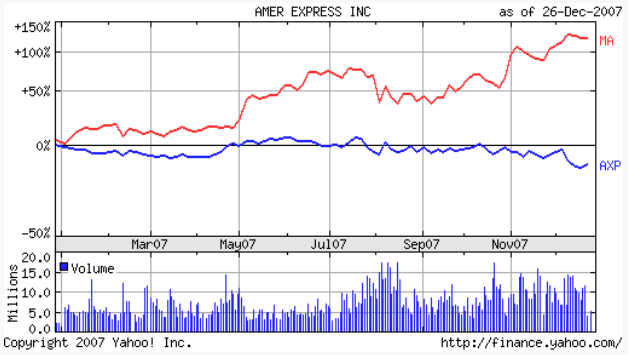
- Design:
  - Different methods for regression in one picture.
  - Different methods for classification in one picture.
  - Model Complexity and Bias-Variance tradeoff.
  - Cover's Theorem

- Assessment:
  - Different measures.
  - Different paradigms.
  - Different estimators.

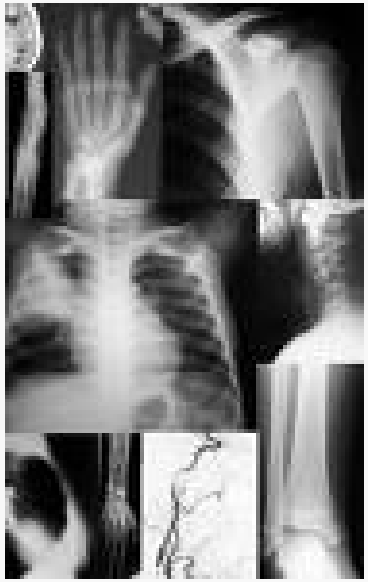
- Concluding Remarks

Diverse Applications and One Theory

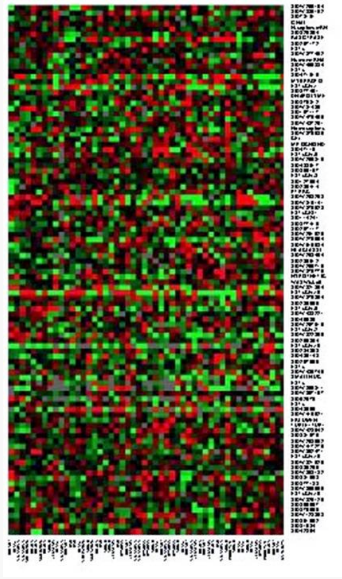
Automatic Target Recognition (ATR)



Stock market prediction



Medical Image Diagnoses



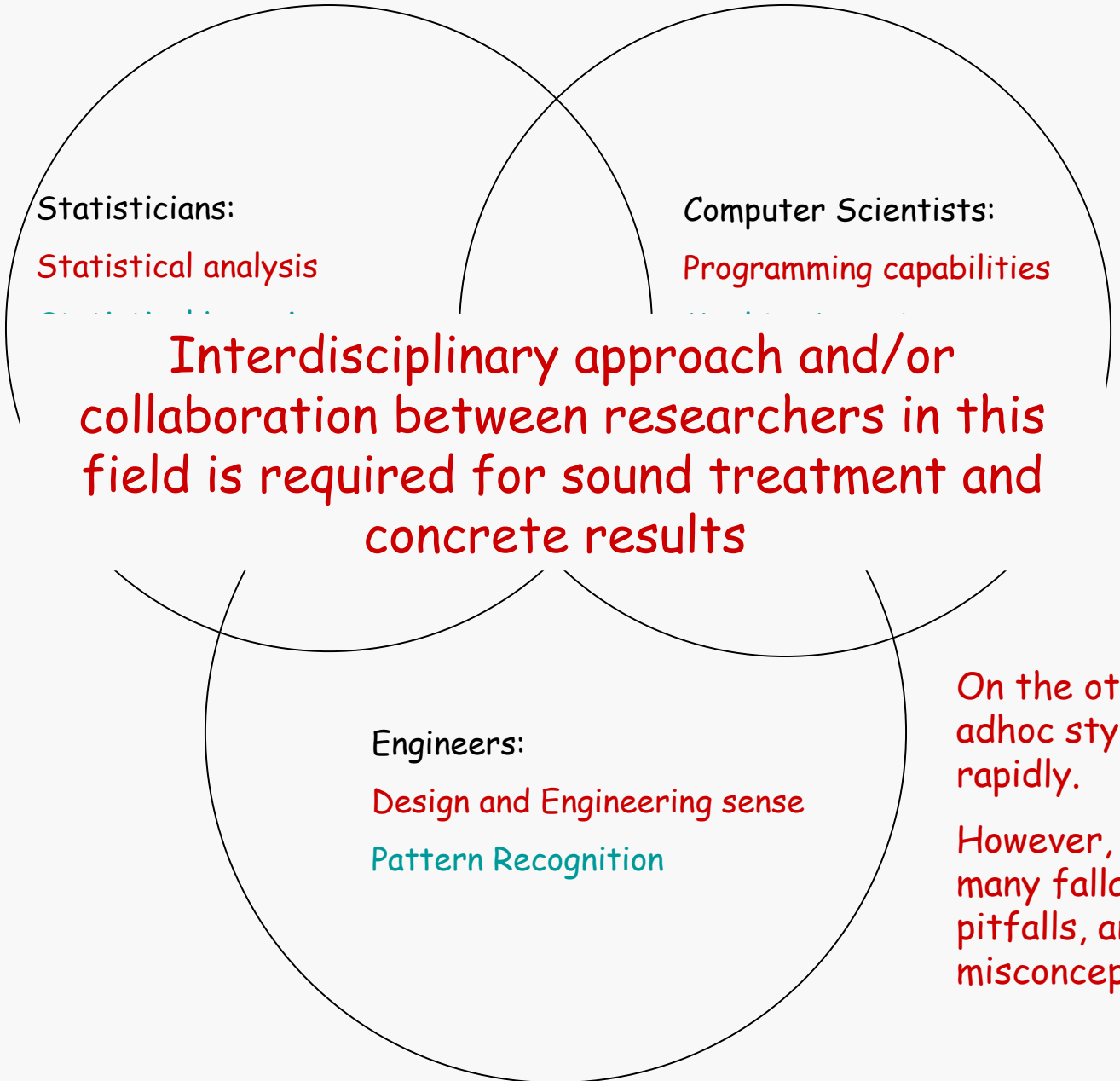
DNA Microarray analysis

Very ill-posed problem, $p >> n$

Who is working in the field?

The mandatory ingredients are Statistics, Probability and mathematical rigor

However, mathematical rigor alone without design talents and trial and error procedures is rigid.



On the other hand, adhoc style invents rapidly.

However, it drives many fallacies, pitfalls, and misconceptions.

10^2 meter

Now,
you can see your whole landscape;
you can easily connect pieces together if you see the big picture.