Note: keeplearning!

- Homework #4 due next week!
- **Complete the Exercise #2: Final Session**
  - **HW Assignment: Complete FP#2 on MeanShift and make your final submission of your code and results (screen shots/short doc report) via Canvas by tomorrow midnight.**

1

Note:

- Fast Prototyping Exercise #3 on LDA classification starts next week.
  - **HW Assignment: Read the reference paper:** https://bidal.sfsu.edu/~kazokada/csc872/PD3.pdf
  - **Download:** https://bidal.sfsu.edu/~kazokada/csc872/DATA/FaceClassification_Data.zip

Linear Discriminant Analysis

2

PF: Machine Learning Framework

CSC 872

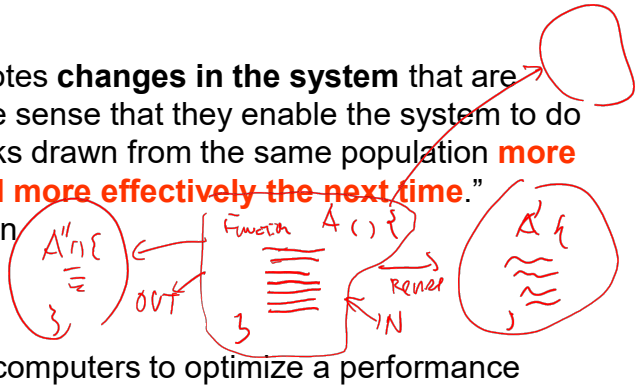
Pattern Analysis and Machine Intelligence

Sources

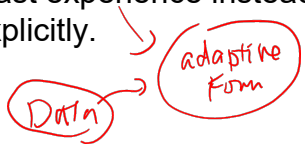
Kevin Murphy, UBC
http://www.cs.ubc.ca/~murphyk/
Andrew Moore, CMU
http://www.cs.cmu.edu/~awm/

What is Machine Learning?

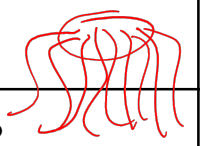
- **For What?**
- "Learning denotes **changes in the system** that are **adaptive** in the sense that they enable the system to do the task or tasks drawn from the same population **more efficiently and more effectively ~~the next~~ time**."
  --Herbert Simon



- **How?**
- Programming computers to optimize a performance criterion using example data or past experience instead of coding algorithmic instructions explicitly.



Why Learning?



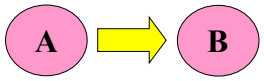
- What are tasks that calls for learning ???
- There is no need to learn to calculate payroll !!!

sort

- **Human expertise does not exist**
  - navigation on Mars
- **Humans are unable to explain their expertise**
  - speech recognition
- **Solution changes in time**
  - computer network routing
- **Solution needs to be adapted to particular cases**
  - user biometrics

Inductive vs Deductive Reasoning

- A entails B
  - Recall our lectures on logic
- Deductive Reasoning
  - **Deriving B as a consequence of A**
  - *All apples are fruit and All fruits grow on trees*
  - *Therefore all apples grow on trees*



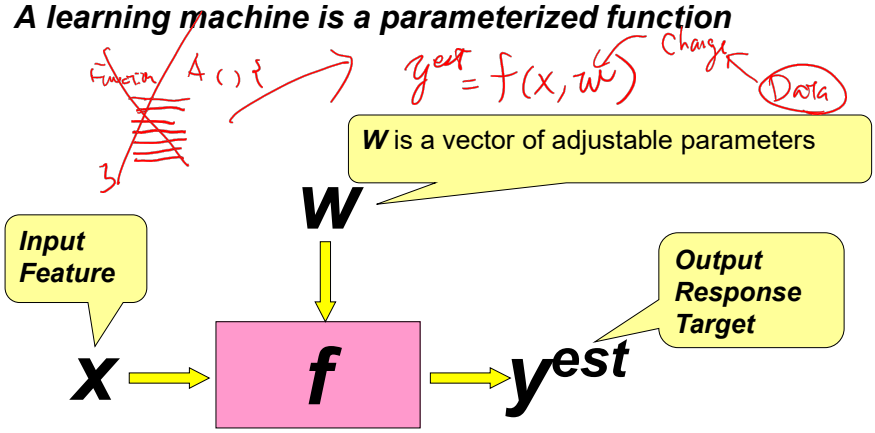
logically invalid!

- Inductive Reasoning ← **THIS IS LEARNING !!!**
  - **Inferring some A from multiple instances of B**
  - *Fuji apples grow on trees and so do McIntosh apples*
  - *Therefore all apples grow on trees*

How likely that all apples grow on trees?

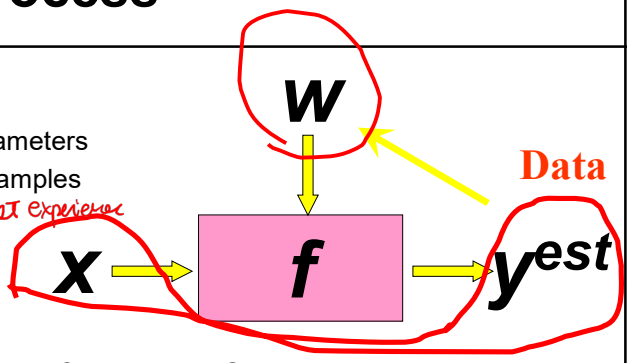
Learning Machine

- A learning machine ***f*** takes an input ***x*** and transforms it, somehow using weights ***w***, into a predicted output y^{est}
- ***A learning machine is a parameterized function***



Learning Process

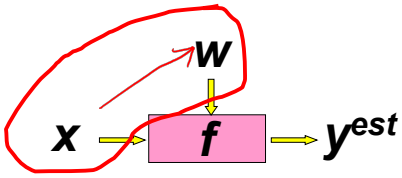
- **Data → *W***
- Deriving a set of parameters of a function from examples



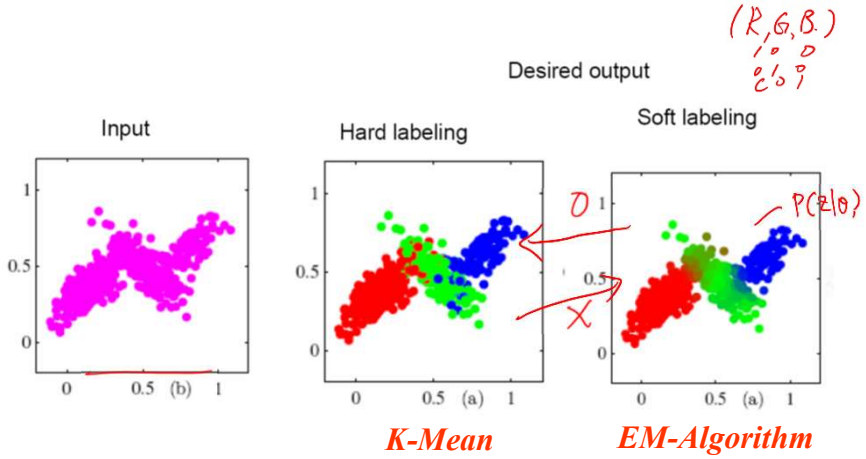
- **What is the nature of your data?**
- **What is the form of your function?**
- **Supervised**
- **Unsupervised**
- **Reinforcement**

PF: Unsupervised Learning

- **No output for training = $\{x_i\}$**
- **Learning *"data structure"***
- ***"what normally happens"***
  - Probabilistic Density Estimation
    - Parametric: GMM, EM
    - Non-Parametric: KDE, MeanShift
- ***"self-similar patterns"***
  - Clustering
  - Dimensionality reduction
    - PCA

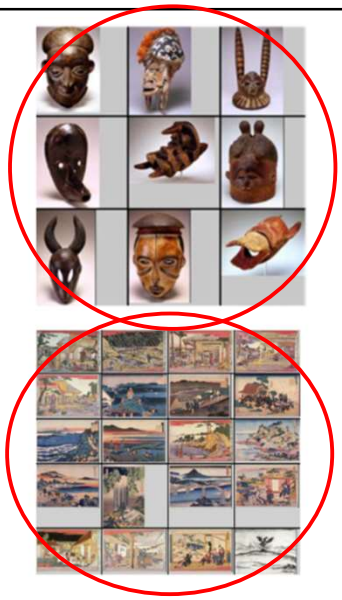


Examples: Clustering



K=3 is the number of clusters, here chosen by hand

Examples: Clustering Arts



Model each cluster by dimensional reduction method like PCA

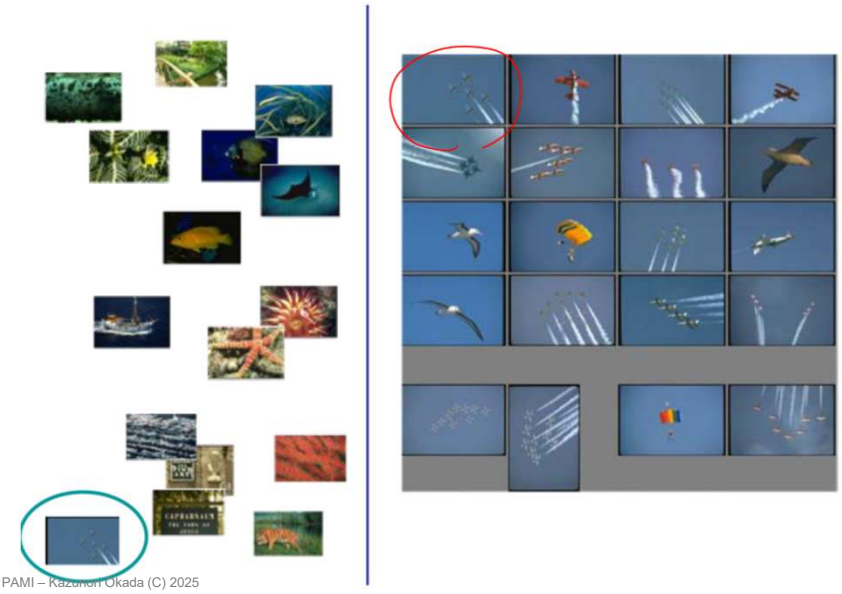
11

Examples: Content-based Image Retrieval

Image Search

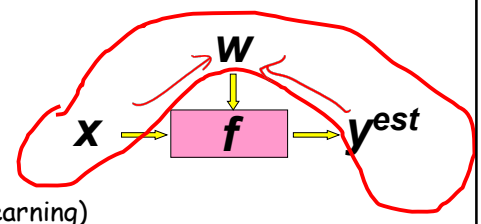
12

Examples: Content-based Image Retrieval



PF: Supervised Learning

- **Training data pairs = $\{(x_i, y_i)\}$**
- **Learning "a mapping from x to y"**
  $\{(x_1,y_1),(x_2,y_2),\ldots\} = \{(x_i,y_i) \mid i=1,\ldots,N\}$
- *Methods differ in terms of*
  - **The form of function** (hypothesis space)
  - **The estimation method** (used to find the best hypothesis)
  - Bayesian Classifier
  - Bayesian Regression
  - Neural Nets
  - Decision Tree/Forest
  - SVM
  - Logistic Regression
  - Boosting (Meta/Ensemble Learning)
  - Convolutional Neural Nets

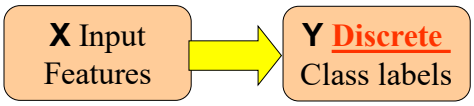


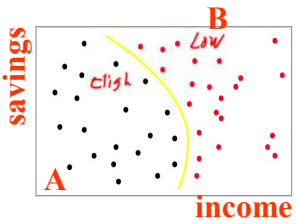
What Supervised Learning good for?

- **Prediction of future cases:**
  - Use the rule to predict the output of future inputs
  - E.g., ***local weather prediction***
- **Knowledge extraction:**
  - Learn the rule that is easy to understand
  - E.g., ***financially low risk if high both in income/savings***
- **Compression:**
  - Learn the rule that is simpler than the data it explains
  - E.g., ***Fed's interest rate increase mortgage rate***
- **Outlier Detection:**
  - Exceptional cases that are not covered by the rule
  - E.g., ***fraud***

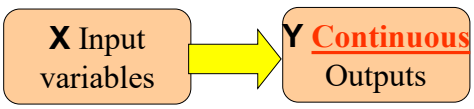
Classification vs Regression

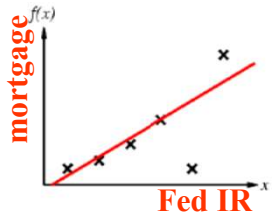
- Nature of output different!!!
- **Classification**





- **Regression**





Generative vs. Discriminative

- Classification by learning discriminant function

$$y^{est} = \hat{f}(\boldsymbol{x})$$

Discriminative

- Bayesian classification

$$y^{est} = \text{argmax}_y p(y|\boldsymbol{x})$$

Generative

– **Discriminative** model

$$p(y|\boldsymbol{x}) = \hat{g}(\boldsymbol{x})$$

Direct

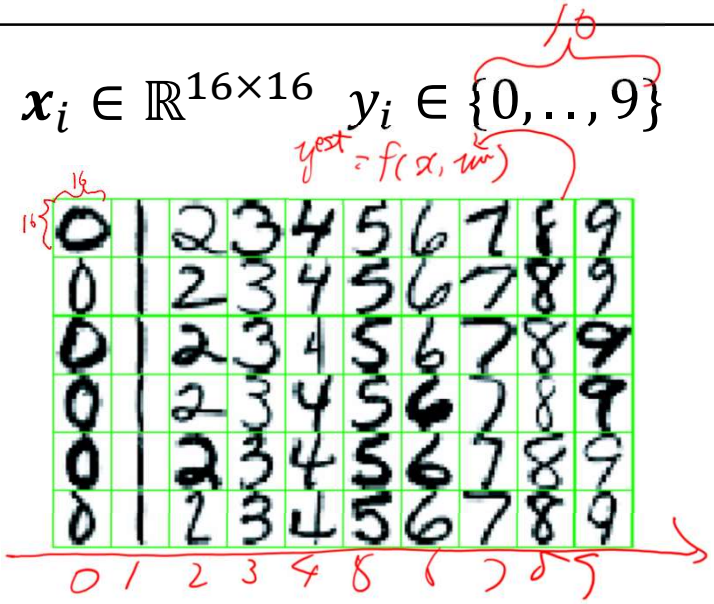
– **Generative** model

$$p(\boldsymbol{x}|y), p(y) \rightarrow \text{Bayes Rule}$$

17

Examples: Handwritten Digit Recog

$$\boldsymbol{x}_i \in \mathbb{R}^{16\times16} \quad y_i \in \{0,..,9\}$$



18

Examples: Face Recognition

Training examples of a person

$\boldsymbol{x}_i \in \mathbb{R}^{128\times128}$

$y_i \in \{Ken, Sue, \dots\}$

Test images

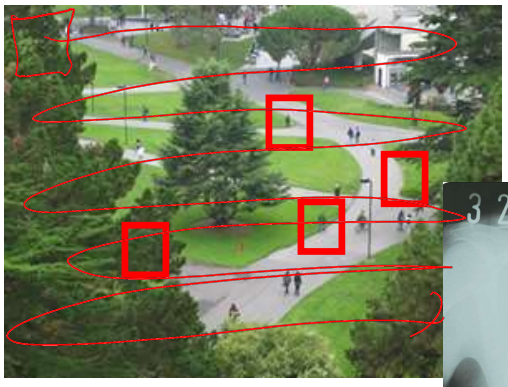
$y^{est} = f(x, w)$

PCA/NMC

AT&T Laboratories, Cambridge UK
http://www.uk.research.att.com/facedatabase.html

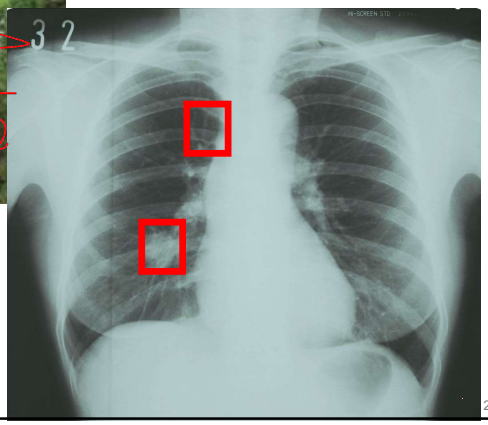
Examples: Object Detection

$y^{est} = f(x, w)$



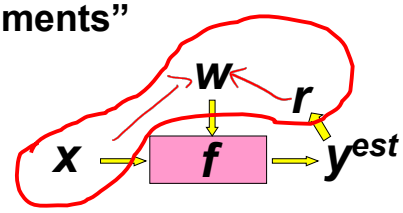
$\boldsymbol{x}_i \in \mathbb{R}^{16\times16}$

$y_i \in \{\text{Yes}, \text{No}\}$



STOP: PF: Reinforcement Learning

- **No output but delayed reward = $\{(x_i, r_i)\}$**
- **Learning "credit assignments"**



- ***E.g., playing chess***
  - ***"which action led me to winning the game?"***
  - ***Reward tells you if an action was good or bad***
  - ***Try to learn f only with such information…***

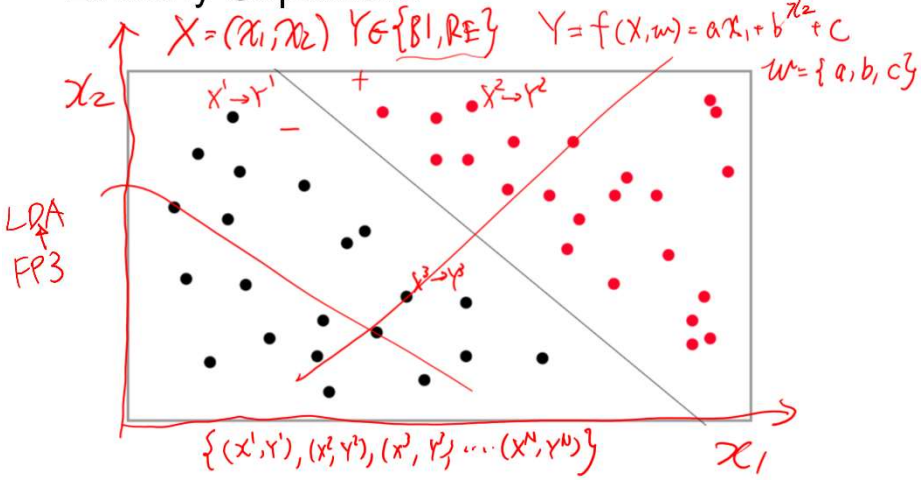
Major Problem Major Problem !!!

- OK. I have this problem at hand
- I want to do machine learning to solve it.
- I learned in this or that course MANY machine learning algorithms…

- **WHICH ONE AM I GONNA USE???**

Well the answer depends on data… Discriminant function example

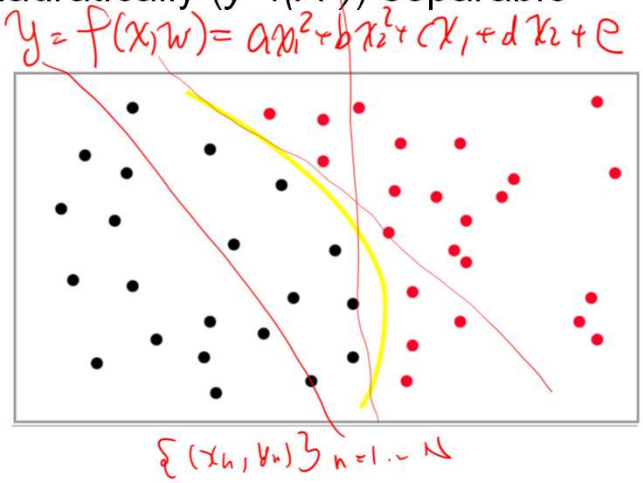
- Linearly Separable



23

Discriminant function

- Quadratically ($y=f(X^2)$) separable



24

Training (Empirical) Errors

Given N training samples $\{(\boldsymbol{x_n}, y_n)\}$

You learned a function $\hat{f}$

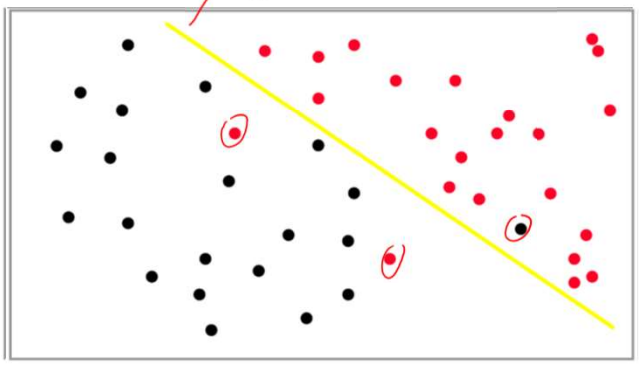
$$err(D) = \frac{1}{N}\sum_{n=1}^{N} I(\hat{f}(\boldsymbol{x_n}) \neq y_n)$$

A learning algorithm is designed to reduce this error

Over-fitting & Under-fitting data

- Noisy Data

Over-fitting: memorize irrelevant details of data



Under-fitting: ignore essential details of data

Regression Example

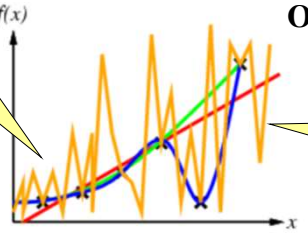
Each learned function is called a **hypothesis *h***. A set of hypothesis form a **hypothesis space H**!

$H = \{R, G, B, Y\}$

$E(R) > E(G) > E(B) = E(Y) = 0$

Or even this!!!

$V \subseteq H$



Version space is formed by hypotheses with **zero** training errors.

$V = \{B, Y\}$

Higher Order Polynomial

These two hypotheses in the version space has the same training error !!!

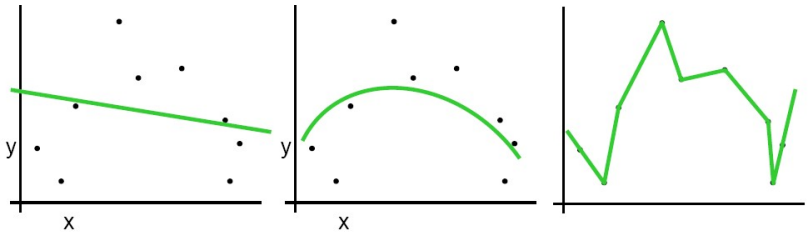
27

No Free Lunch Theorem

- ***Any hypothesis that agrees with all your data is as good as any other on average***
  - Wolpert and Macready (1995)
  - Wolpert (2001)
- You have two machine learning algorithms
- Ask which one is better in general?
- NFL: on average over all problems, they are the same!
- **No general take-it-all solution!**

28

Choose the best given data?



Why not choose the method with the best fit to the data?

Empirical Risk Minimization: find a hypothesis that minimizes the training error

NFL tells us that **it is not a good idea !**

Model Selection

- Problem:

  Choose the best model/algorithm given a task/data
  - Linear, Quadratic, Polynomial functions?
  - Generative vs Discriminative?
  - CNN vs SVM vs AdaBoost vs Random Forest?
- Learning is an ill-posed problem; data is not sufficient to find a unique solution
- We need an *inductive bias / assumptions / heuristics* about the hypothesis space …
- AN OPEN PROBLEM!

What-to-do? #1 Ockham's Razor

- Bias to prefer the simplest hypothesis consistent with data!
- Why?
- A simpler hypothesis is less likely to be correct "by chance" and is therefore more likely to be correct on unseen new data (generalization)
- Other reasons?
  - Simpler to use (lower computational complexity)
  - Easier to train (lower samplecomplexity)
  - Easier to explain (more interpretable) LDA

Generalization and Test Errors

Given N training samples $\{(\boldsymbol{x_n}, y_n)|n = 1, \ldots N\}$ S
You learned a function $\hat{f}$ Test Set LS

- Generalization Errors: average error for ALL possible data in domain D (cannot calculate)

$$E[err(D)] = \sum_{\boldsymbol{x},y \in D} p(\boldsymbol{x}, y) I(\hat{f}(\boldsymbol{x}) \neq y)$$

- Test Errors: approximate generalization error…

$$\widehat{err} = \frac{1}{\#\text{TestSet}} \sum_{k \in \text{TestS}} I(\hat{f}(x) \neq y) \quad \subset D$$

not used in Training

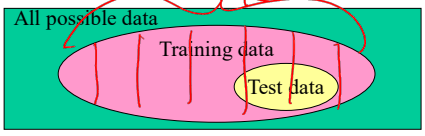
Prepare & use different data set

Model Selection using Test Data

- **Goal**: *Choose the model that generalize best*
- **Generalization**: *How well a model performs on new data?*
- Steps:
  - Prepare "test data" that is different from your "training data"
  - Compute "test error" for each model (approximating "generalization error")
  - Pick the model with lowest "test error"
- **Question: how to prepare the test data?**

Cross Validation (CV)

- **We hold back a fraction of the training data, called a validation set, and measure performances on that using test error.**



- **Random Test Set CV**
  - Randomly sample X% of data as a validation set & repeat to get an average error.
- **Leave-One-Out CV**
  - Validation set contains only one data point & repeat N times
- **K-fold CV**
  - Data into K partitions then repeat this K times

More Model Selection Criteria

- Can we do model selection without doing test data sampling and CV???
- YES! These only uses training errors

- **AIC**: (Akaike information criterion)
- **BIC**: (Bayesian information criterion)
- **SRMVC**: (structural risk minimization with VC-dimension)

Which criterion should I use?

- CV: high variance due to sampling…
- SRMVC, AIC, BIC only uses training errors
- SRMVC wildly conservative
- AIC: asymptotically same as LOO-CV
- BIC: asymptotically similar to K-fold CV
- BIC: best structure instead of best predictor

- **Open problem!!!**

Summary

- Machine Learning Framework
  - Unsupervised/Supervised/Reinforcement Learning
  - Classification vs Regression
  - Generative vs Discriminative
  - Model Selection: Occam's Razor, Cross-Validation
- Next
  - Learning for Classification
  - Bayesian Classification
  - Discriminant Analysis
  - Performance Measure
  - FP#3 on Linear Discriminant Analysis: **READ THE PAPER!!!**