Statistics and Machine Learning Fall, 2005

鮑興國 and 李育杰

National Taiwan University of Science and Technology

General Information:

- Instructor: 李育杰、吳漢銘、陳素雲、陳君厚、張源俊、鮑興國
- Textbook:
  - Introduction to Machine Learning, E. Alpaydin, MIT, 2004.
- Lectures
  - Time: Friday 14:00~15:15; 15:30~16:45
- Grading
  - Homeworks (40%), Exam (30%), Final Project (30%)
- Course page: http://www3.stat.sinica.edu.tw/stat2005w/schedule.htm

Mathematical Background You Will Need in the Class

- Mutli-Variable Calculus
  - What is the *gradient* of a differentiable function?
  - What is the *Hessian* of a twice differentiable function?
- Linear Algebra
  - How to compute the distance between two parallel hyperplanes in R^n?
  - Eigenvalue, positive definite matrix, inner product, projection matrix etc.
- Probability
  - Random variables, probability distributions, conditional probability, Bayes' rule, expected value, variance etc.
- Statistics
  - Testing hypothesis, confidence interval etc.

Course Overview

- Introduction & Supervised Learning
- Optimization (MATLAB)
- Introduction to R and WEKA & Final project list
- Clustering (Unsupervised Learning)
- Parametric Methods and Multivariate Methods
- Bayesian Decision Theory & Bayesian Network
- Linear Discrimination
- Dimensionality Reduction
- Nonparametric Methods 1 & 2
- Multilayer Perceptrons
- Support Vector Machines
- Homework Discussion
- Data Visualization

Software Packages & Datasets

- MLC++
  - Machine learning library in C++
  - http://www.sgi.com/tech/mlc/
- WEKA
  - http://www.cs.waikato.ac.nz/ml/weka/
- Statlib
  - Data, software and news from the statistics community
  - http://lib.stat.cmu.edu
- GALIB
  - MIT GALib in C++
  - http://lancet.mit.edu/ga
- Delve
  - Data for Evaluating Learning in Valid Experiments
  - http://www.cs.utoronto.ca/~delve
- UCI
  - Machine Learning Data Repository UC Irvine
  - http://www.ics.uci.edu/~mlearn/MLRepository.html
- UCI KDD Archive
  - http://kdd.ics.uci.edu/summary.data.application.html

Major conferences in ML

- ICML (International Conference on Machine Learning)
- ECML (European Conference on Machine Learning)
- UAI (Uncertainty in Artificial Intelligence)
- NIPS (Neural Information Processing Systems)
- COLT (Computational Learning Theory)
- IJCAI (International Joint Conference on Artificial Intelligence)
- MLSS (Machine Learning Summer School)

Three Big Events of Statistics & ML in Taiwan, 2006

- Statistics and ML winter camp (January 16~20)
  - Advanced topics in ML
  - Limited 20 ~25 students
  - Supported by Institute of Statistics Science
- Machine Learning Summer School will be held in Taiwan (July, 23~August, 2)
- International Workshop on Statistics and ML will be right after MLSS (August 2 ~August 4)

What is Learning All about?

- Get knowledge of by study, experience, or be taught
- Become aware by information or from observation
- Commit to memory
- Be informed of or receive instruction

A Possible Definition of Learning

- Things learn when they change their behavior in a way that makes them *perform* better in the future.
- Have your shoes *learned* the shape of your foot ?
- In learning the purpose is the learner's, whereas in training it is the teacher's.

Learning & Adaptation

- Machine Learning: 機器學習?
  - Machine → Automatic
  - Learning → Performance is improved
- "A learning machine, broadly defined is any device whose actions are influenced by past experiences." (Nilsson 1965)
- "Any change in a system that allows it to perform better the second time on repetition of the same task or on another task drawn from the same population." (Simon 1983)
- "An improvement in information processing ability that results from information processing activity." (Tanimoto 1990)

Applications of ML

- Learning to recognize spoken words
  - SPHINX (Lee 1989)
- Learning to drive an autonomous vehicle
  - ALVINN (Pomerleau 1989)
  - Taxi driver vs. Pilot
- Learning to pick patterns of terrorist action
- Learning to classify celestial objects
  - (Fayyad et al 1995)
- Learning to play chess
  - Learning to play go game (Shih, 1989)
  - Learning to play world-class backgammon (TD-GAMMON, Tesauro 1992)
- Information Security: Intrusion detection system (normal vs. abnormal)
- Bioinformation

Prediction is the Key in ML

- We make predictions all the time but rarely investigate the processes underlying our predictions.
- In carrying out scientific research we are also governed by how theories are evaluated.
- To automate the process of making predictions we need to understand *in addition* how we search and refine “theories”

Types of learning problems

- A rough (and somewhat outdated) classification of learning problems:
  - Supervised learning, where we get a set of training inputs and outputs
    - classification, regression
  - Unsupervised learning, where we are interested in capturing inherent organization in the data
    - clustering, density estimation
  - Semi-supervised learning
  - Reinforcement learning, where we only get feedback in the form of how well we are doing (not what we should be doing)
    - planning

Issues in Machine Learning

- What algorithms can approximate functions well and when?
- How does the number of training examples influence accuracy?
- How does the complexity of hypothesis representation impact it?
- How does noisy data influence accuracy?
- What are the theoretical limits of learnability?

Learning a Class from Examples: Inductive (歸納)

- Suppose we want to learn a class C
  - Example: “sports car”
  - Given a collection of cars, have people label them as sports car (positive example) or non-sports car (negative example)
  - Task: find a description (rule) that is shared by all of the positive examples and none of the negative examples
  - Once we have this definition for C, we can
    - predict – given a new unlabeled car, predict whether or not it is a sports car
    - describe/compress – understand what people expect in a car

Choosing an Input Representation

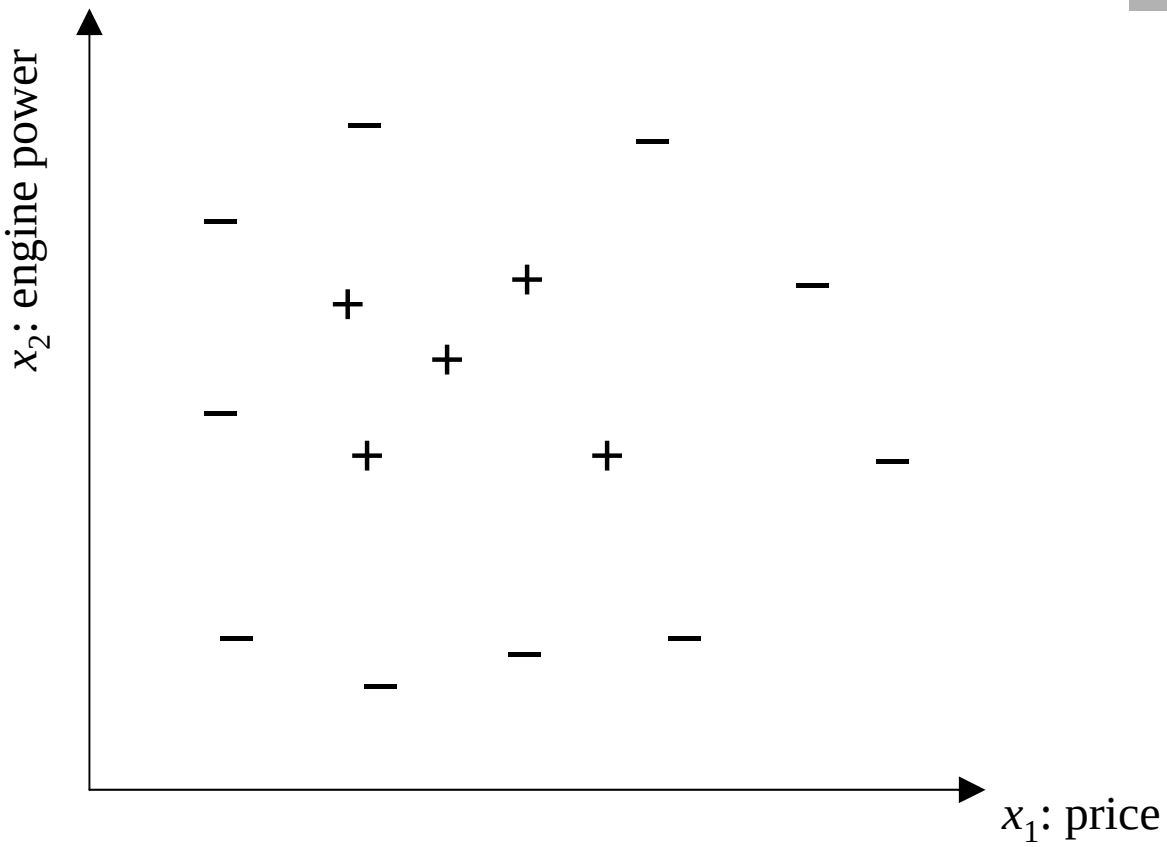
- Suppose that of all the features describing cars, we choose price and engine power. Choosing just two features
  - makes things simpler
  - allows us to ignore irrelevant attributes
- Let
  - x_1 represent the price (in USD)
  - x_2 represent the engine volume (in cm^3)
- Then each car is represented

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

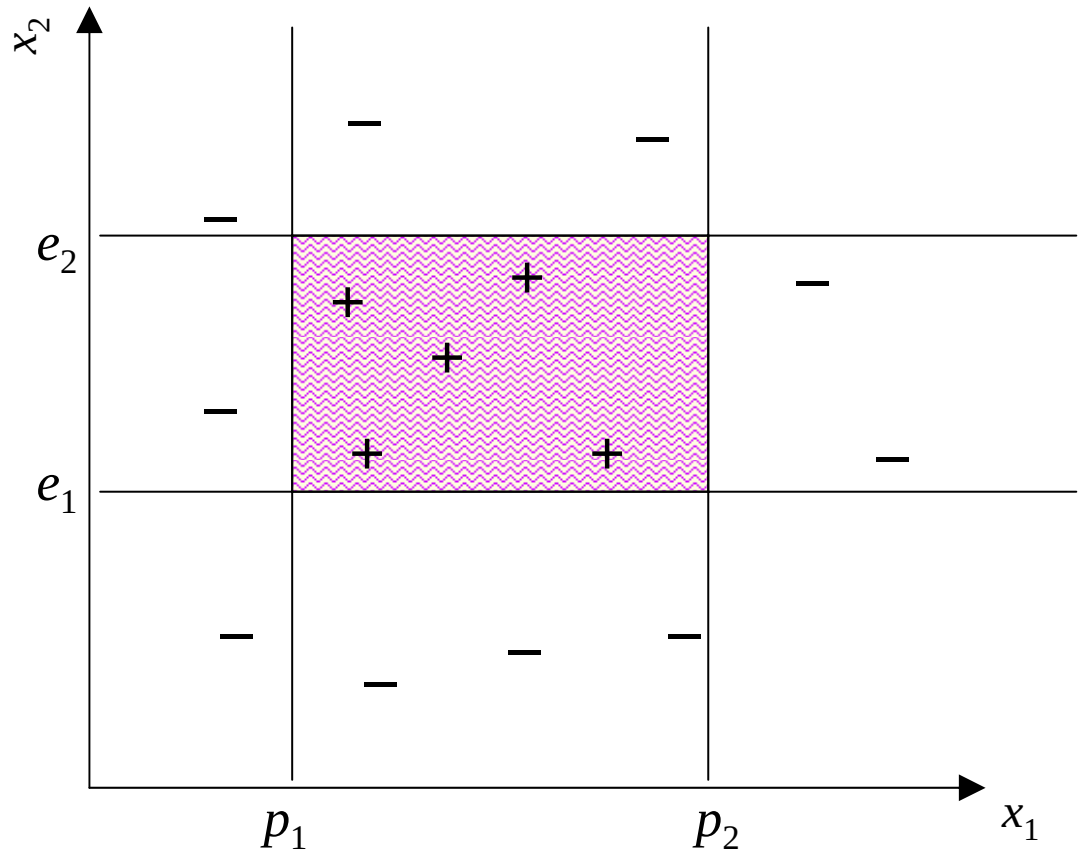
- and its label y denotes its type $y = \begin{cases} 1 & \text{if } \mathbf{x} \text{ is a positive example} \\ 0 & \text{if } \mathbf{x} \text{ is a negative example} \end{cases}$
- each example is represented by the pair $(\mathbf{x}, y)$
- and a training set containing N examples is represented by

$$\mathcal{X} = \{\mathbf{x}^t, y^t\}_{t=1}^{N}$$

Plotting the Training Data



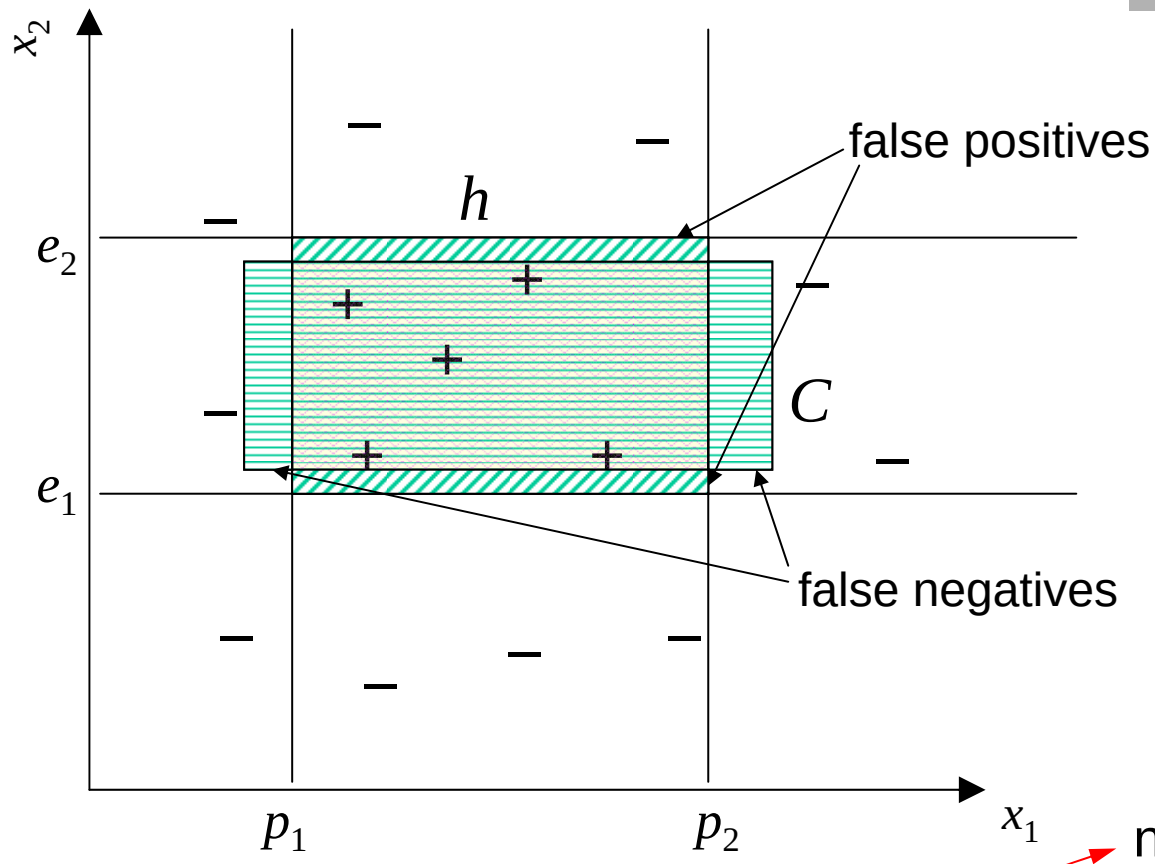
Hypothesis Class



suppose that we think that for a car to be a sports car, its price and its engine power should be in a certain range:

$$(p_1 \leq \text{price} \leq p_2) \text{ AND } (e_1 \leq \text{engine} \leq e_2)$$

Concept Class



no training errors

suppose that the actual class is C

task: find $h \in \mathcal{H}$ that is **consistent** with $\mathcal{X}$

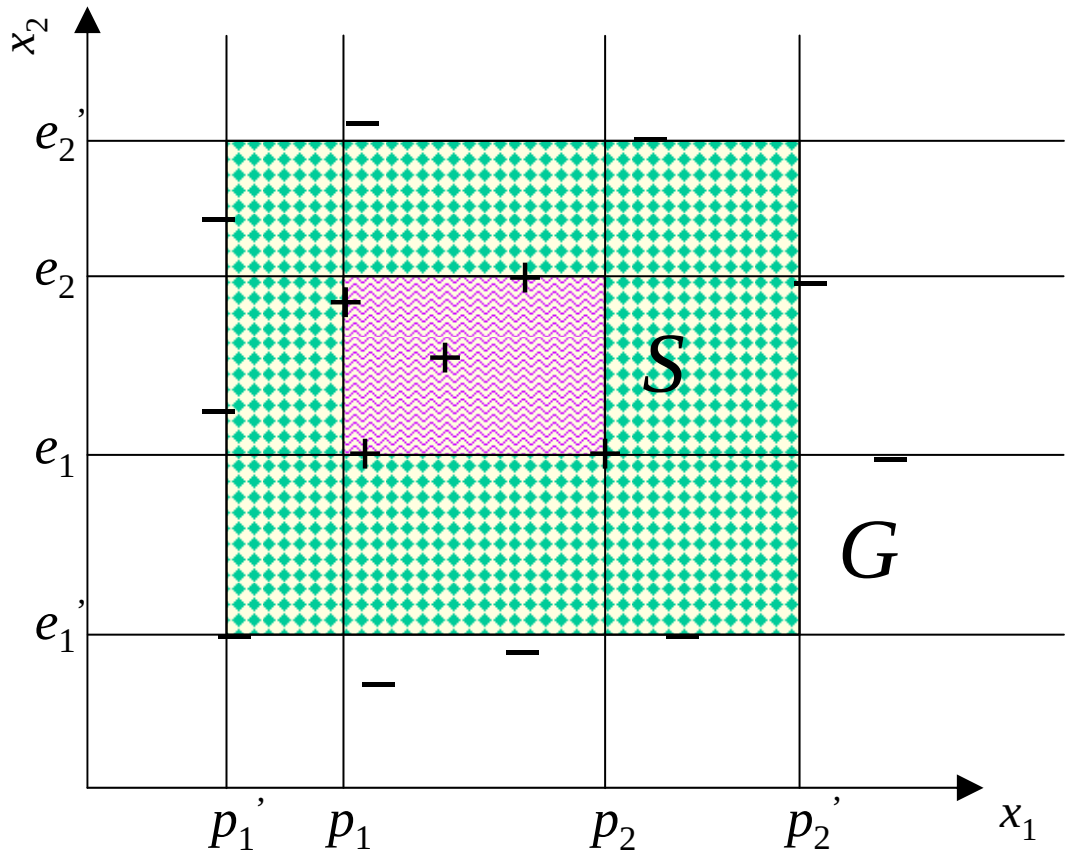
Choosing a Hypothesis

- Empirical Error: proportion of training instances where predictions of h do not match the training set

$$E(h|X) = \sum_{t=1}^{N} 1(h(\mathbf{x}^t) \neq y^t)$$

- Each (p_1, p_2, e_1, e_2) defines a hypothesis $h \in \mathcal{H}$
- We need to find the best one…

Hypothesis Choice

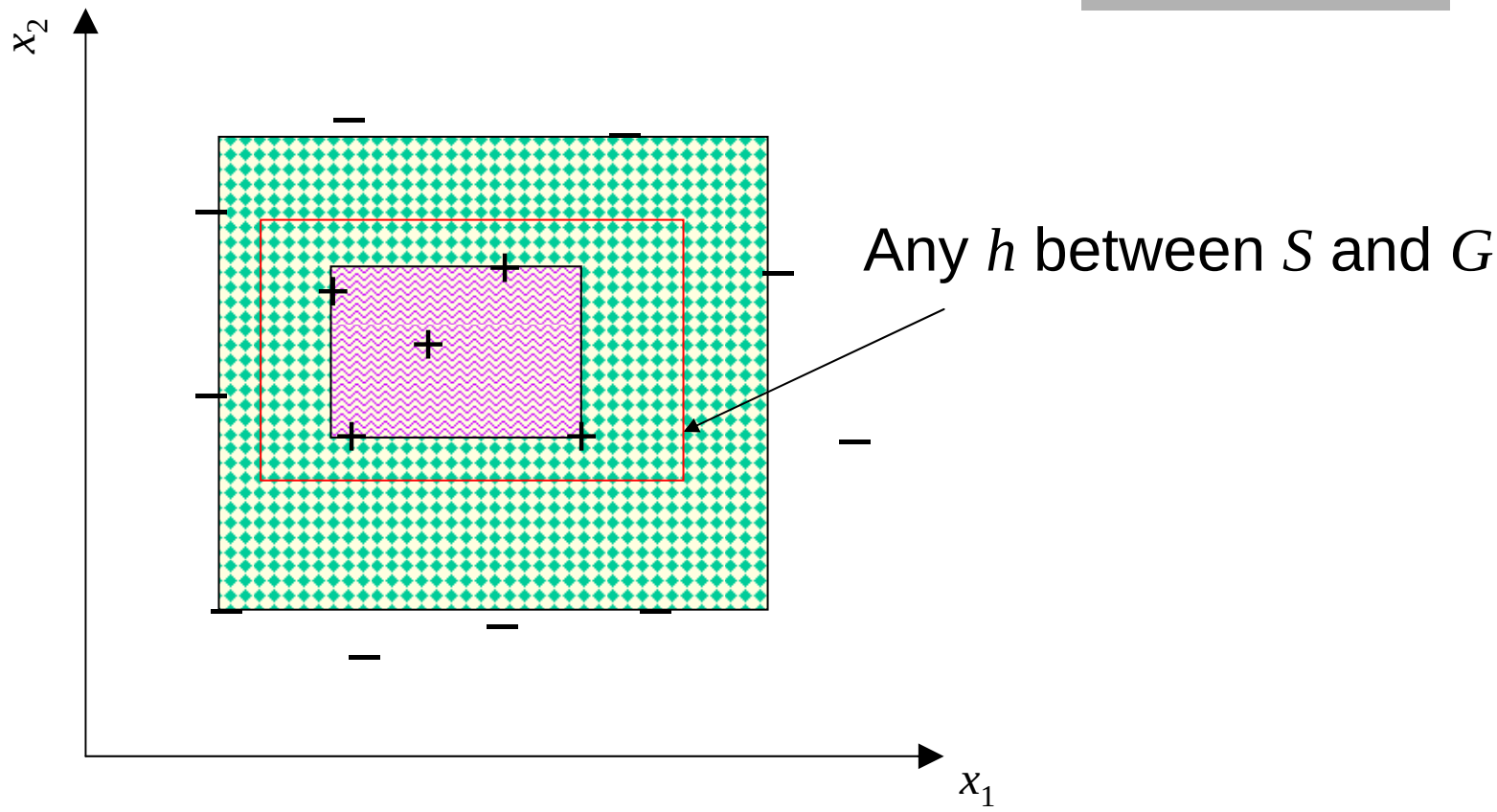


Most specific hypothesis S
Most general hypothesis G

Most specific?

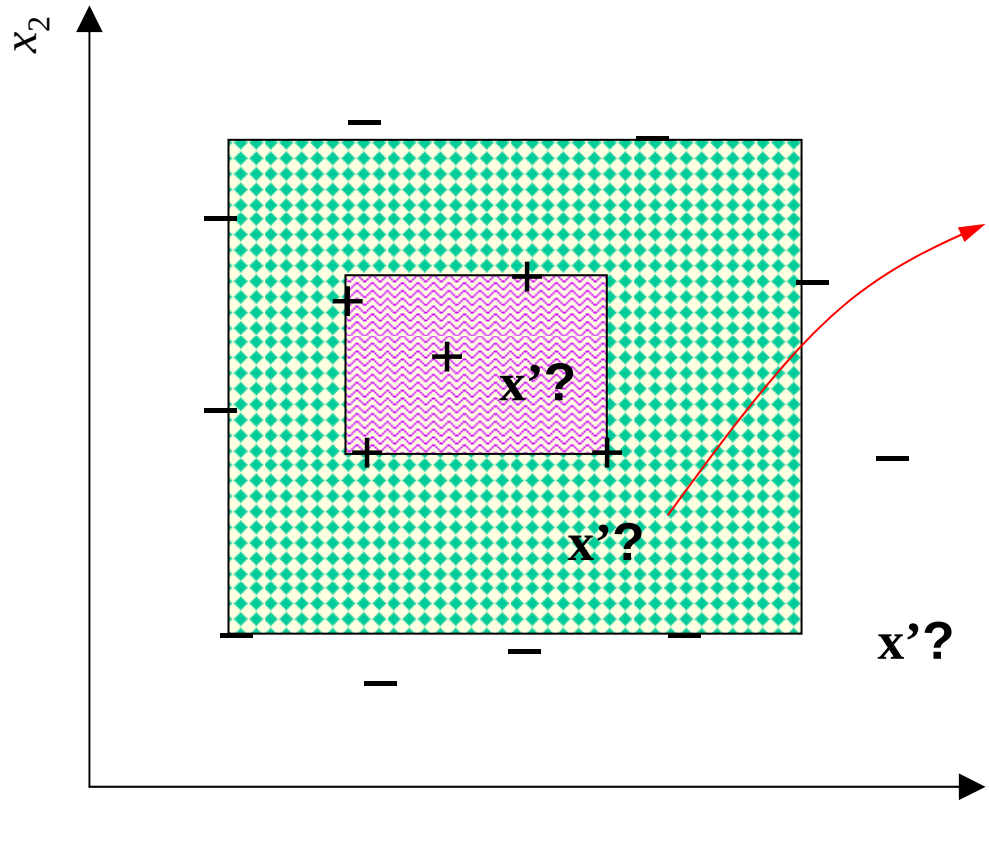
Most general?

Consistent Hypothesis



G and S define the boundaries of the Version Space.
The set of hypotheses more general than S and more specific than G forms the **Version Space**, the set of consistent hypotheses

Now what?



- *Using the average of S and G or just rejecting it to experts?*

How do we make prediction for a new **x'**?

VS on another Example

example #	x_1	x_2	x_3	x_4	y
1	1	1	0	0	1
2	1	0	0	0	1
3	0	1	1	1	0

- $\mathcal{H}$ = conjunctive rules

  $S = x_1 \wedge (\neg\, x_3) \wedge (\neg\, x_4)$

  $G = x_1\,,\ \neg\, x_3\,,\ \neg\, x_4$

Issues

- Hypothesis space must be flexible enough to represent concept
- Making sure that the gap of S and G sets do not get too large
- Assumes no noise!
  - inconsistently labeled examples will cause the version space to **collapse**
  - there have been extensions to handle this…

Goal of Learning Algorithms

- The early learning algorithms were designed to find such an accurate fit to the data.
- The ability of a classifier to correctly classify data *not in the training set* is known as its *generalization.*
- Bible code? 1994 Taipei Mayor election?
- Predict the real future *NOT fitting the data in your hand or predict the desired results*

Binary Classification Problem

Learn a Classifier from the Training Set

Given a training dataset

$$S = \{(x^i, y_i) \big| x^i \in R^n, y_i \in \{-1, 1\}, i = 1, \ldots, m\}$$

$$x^i \in A_+ \Leftrightarrow y_i = 1 \;\; \& \quad x^i \in A_- \Leftrightarrow y_i = -1$$

Main goal: Predict the unseen class label for new data

(I) Find a function $f : R^n \rightarrow R$ by learning from data

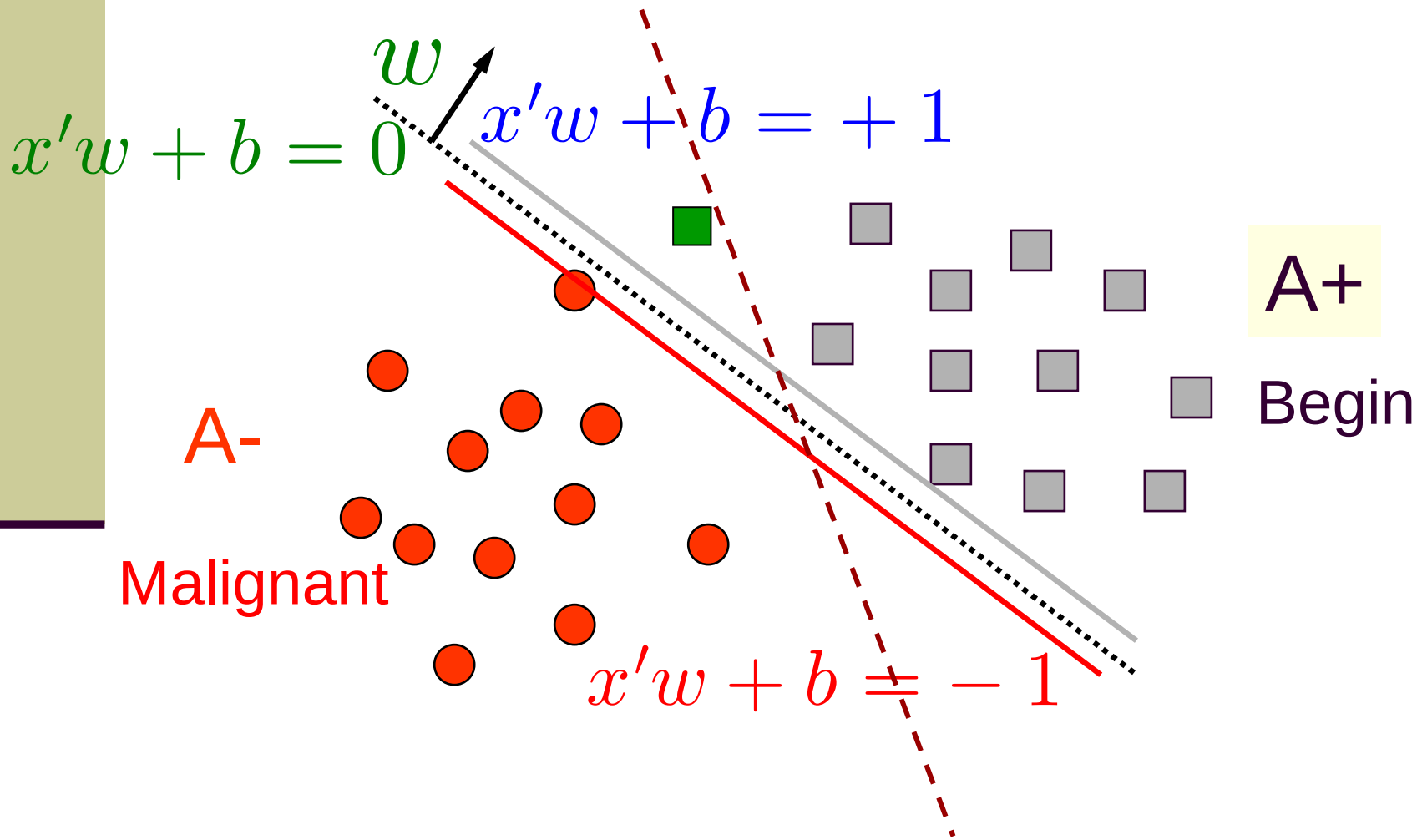
$$f(x) \geqslant 0 \Rightarrow x \in A_+ \;\; and \quad f(x) < 0 \Rightarrow x \in A_-$$

(II) Estimate the *posteriori probability* of label

$$Pr(y = 1|x) > Pr(y = -1|x) \Rightarrow x \in A_+$$

Binary Classification Problem

Linearly Separable Case



Probably Approximately Correct Learning
pac Model

- Key assumption:

  Training and testing data are generated *i.i.d.* according to a *fixed but unknown* distribution $\mathcal{D}$

- Evaluate the "quality" of a hypothesis (classifier) $h \in H$ should take the *unknown* distribution $\mathcal{D}$ into account (*i.e.* "average error" or "expected error" made by the $h \in H$)

- We call such measure risk functional and denote it as $err_{\mathcal{D}}(h) = \mathcal{D}\{(x, y) \in X \times \{1, -1\} | \ h(x) \neq y\}$

Generalization Error of pac Model

- Let $S = \{(x^1, y_1), \ldots, (x^l, y_l)\}$ be a set of l training examples chosen i.i.d. according to $\mathcal{D}$
- Treat the generalization error $err_{\mathcal{D}}(h_S)$ as a *r.v.* depending on the random selection of S
- Find a bound of the trail of the distribution of *r.v.* $err_{\mathcal{D}}(h_S)$ in the form $\varepsilon = \varepsilon(l, H, \delta)$
- $\varepsilon = \varepsilon(l, H, \delta)$ is a function of l, H and δ,where $1 - \delta$ is the confidence level of the error bound which is given by learner

Probably Approximately Correct

- We assert:

$$Pr(\{\, err_{\mathcal{D}}(h_S) > \varepsilon = \varepsilon(l, H, \delta)\}) < \delta$$

or

$$Pr(\{\, err_{\mathcal{D}}(h_S) \leqslant \varepsilon = \varepsilon(l, H, \delta)\}) \geqslant 1 - \delta$$

- The error made by the hypothesis h_s will be less then the error bound $\varepsilon(l, H, \delta)$ that is not depend on the unknown distribution $\mathcal{D}$

Probably Approximately Correct Learning

- We allow our algorithms to fail with probability δ.
- **Finding an approximately correct hypothesis with high probability**

  Imagine drawing a sample of N examples, running the learning algorithm, and obtaining h. Sometimes the sample will be unrepresentative, so we want to insist that 1 – δ the time, the hypothesis will have error less than ε.

$$Pr(\{err_{\mathcal{D}}(h_S) > \varepsilon = \varepsilon(N, H, \delta)\}) < \delta$$

- For example, we might want to obtain a 99% accurate hypothesis 90% of the time.

PAC *vs.* 民意調查

- 成功樣本爲1265個，以單純隨機抽樣方式（SRS）估計抽樣誤差，在95%的信心水準下，其最大誤差應不超過±2.76%。

$$Pr(\{\, err_{\mathcal{D}}(h_S) \leqslant \varepsilon = \varepsilon(l, H, \delta)\}) \geqslant 1 - \delta$$

$$l = 1265, \;\; \varepsilon(l, H, \delta) = 0.0276, \;\; \delta = 0.05$$

Find the Hypothesis with Minimum Expected Risk?

- Let $S = \{(x^1, y_1), \ldots, (x^l, y_l)\} \subseteq X \times \{-1, 1\}$ be the training examples chosen i.i.d. according to $\mathcal{D}$ with the probability density $p(x, y)$
- The expected misclassification error made by $h \in H$ is
$$R[h] = \int_{X \times \{-1,1\}} \frac{1}{2}|h(x) - y| dp(x, y)$$
- The *ideal* hypothesis h^*_{opt} should has the smallest expected risk $R[h^*_{opt}] \leqslant R[h], \quad \forall h \in H$

Unrealistic !!!

Empirical Risk Minimization (ERM)

($\mathcal{D}$ and $p(x,y)$ are not needed)

- Replace the expected risk over $p(x,y)$ by an average over the training example
- The empirical risk: $R_{emp}[h] = \frac{1}{l}\sum_{i=1}^{l}\frac{1}{2}|h(x^i) - y_i|$
- Find the hypothesis h^*_{emp} with the smallest empirical risk $\quad R_{emp}[h^*_{emp}] \leqslant R_{emp}[h], \quad \forall h \in H$
- Only focusing on empirical risk will cause *overfitting*

VC Confidence

(The Bound between $R_{emp}[h]$ & $R[h]$)

- The following inequality will be held with probability $1-\delta$

$$R[h] \leqslant R_{emp}[h] + \sqrt{\frac{v(log(2l/v)+1)-log(\delta/4)}{l}}$$

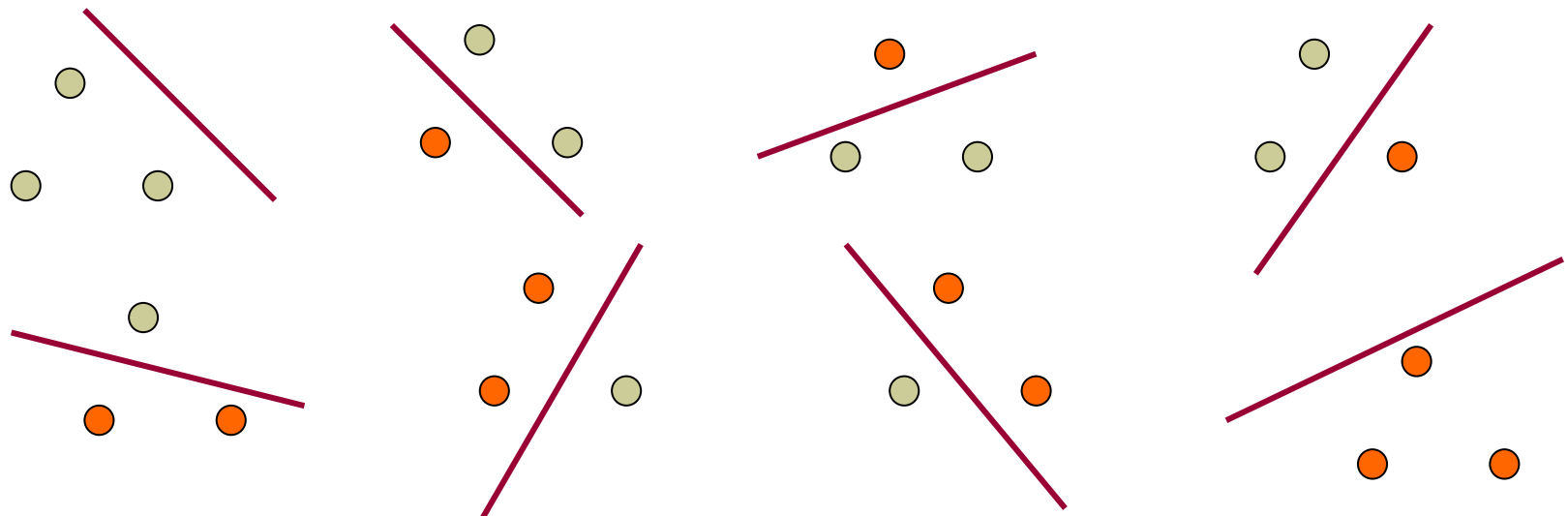
C. J. C. Burges, *A tutorial on support vector machines for pattern recognition*,
Data Mining and Knowledge Discovery 2 (2) (1998), p.121-167

Why We Maximize the Margin? (Based on Statistical Learning Theory)

- The Structural Risk Minimization (SRM):
  - The expected risk will be less than or equal to empirical risk (training error)+ VC (error) bound
- $\|w\|_2 \propto VC\ \ bound$
- $\min VC\ \ bound \Leftrightarrow \min \frac{1}{2}\|w\|_2^2 \Leftrightarrow \max Margin$

Capacity (Complexity) of Hypothesis Space H :VC-dimension

- A given training set S is *shattered* by H if and only if for every labeling of S, $\exists\ h \in H$ consistent with this labeling
- Three (linear independent) points shattered by a hyperplanes in R^2



Shattering Points with Hyperplanes in R^n

Can you always shatter three points with a line in R^2?

Theorem: Consider some set of m points in R^n. Choose a point as origin. Then the m points can be shattered by oriented hyperplanes if and only if the position vectors of the rest points are linearly independent.

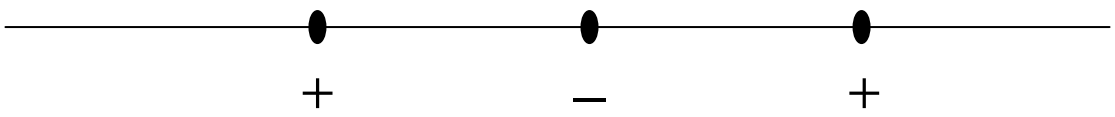
Definition of VC-dimension

(A Capacity Measure of Hypothesis Space H)

- The *Vapnik-Chervonenkis* dimension, $VC(H)$, of hypothesis space H defined over the input space X is the size of the (existent) largest finite subset of X shattered by H
- If arbitrary large finite set of X can be shattered by H, then $VC(H) \equiv \infty$
- Let $H = \{all \ \ hyperplanes \ \ in \ \ R^n\}$ then $VC(H) = n + 1$

Example I

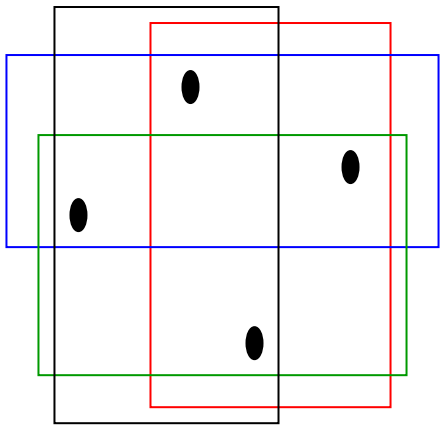
- $x \in \mathbf{R}$, $\mathcal{H}$ = interval on line
  - There exists two points that can be shattered
  - No set of three points can be shattered
  - $\mathrm{VC}(\mathcal{H}) = 2$



  - An example of three points (and a labeling) that cannot be shattered

Example II

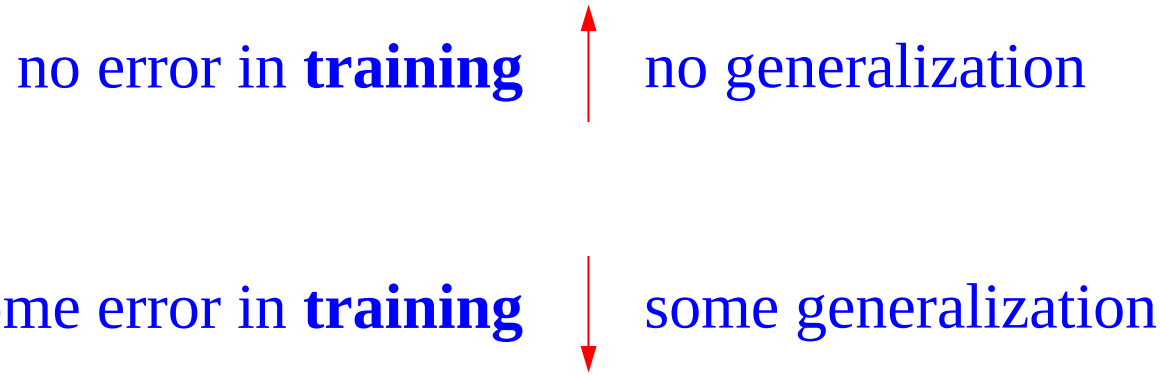
- $\mathbf{x} \in \mathbf{R} \times \mathbf{R}$, $\mathcal{H}$ = Axis parallel rectangles
  - There exist four points that can be shattered
  - No set of five points can be shattered
  - $\mathrm{VC}(\mathcal{H}) = 4$



- Hypotheses consistent with all ways of labeling three positive;
- Check that there hypothesis for all ways of labeling one, two or four points positive

Example III

- A lookup table has infinite VC dimension!



- A hypothesis space with low VC dimension

Comments

- VC dimension is distribution-free; it is independent of the probability distribution from which the instances are drawn
- In this sense, it gives us a worse case complexity (pessimistic)
  - In real life, the world is smoothly changing, instances close by most of the time have the same labels, no worry about *all possible labelings*
- However, this is still useful for providing bounds, such as the sample complexity of a hypothesis class.
- In general, we will see that there is a connection between the VC dimension (which we would like to minimize) and the error on the training set (empirical risk)

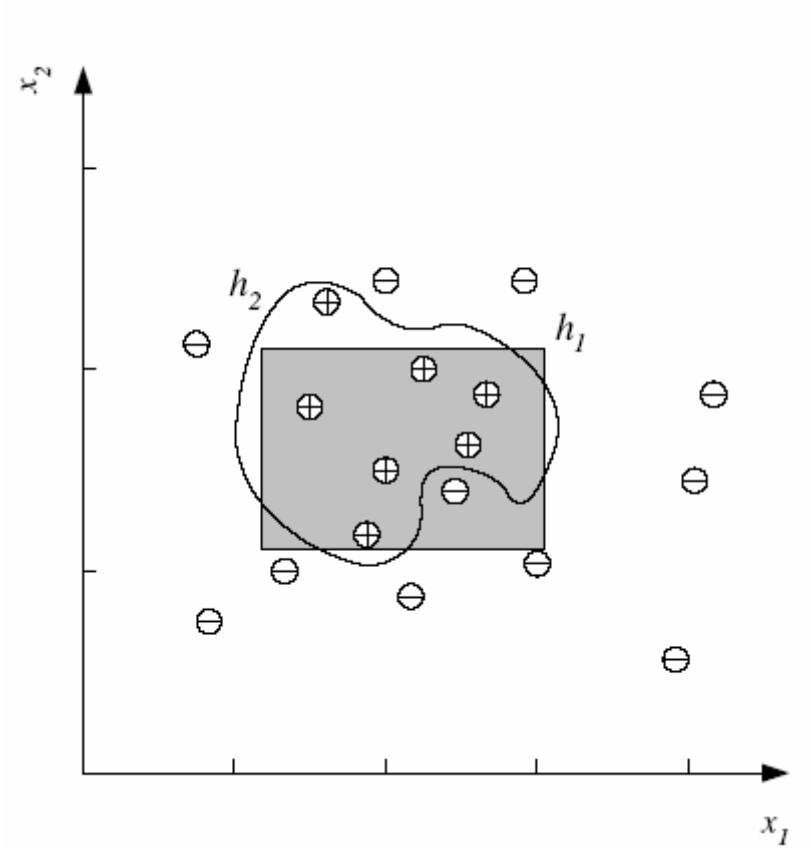
Summary: Learning Theory

- The complexity of a hypothesis space is measured by the VC-dimension
- There is a tradeoff between ε, δ and N

Noise

- Noise: unwanted anomaly in the data
- Another reason we can't always have a perfect hypothesis
  - error in sensor readings for input
  - teacher noise: error in labeling the data
  - additional attributes which we have not taken into account. These are called **hidden** or **latent** because they are unobserved.

When there is noise…



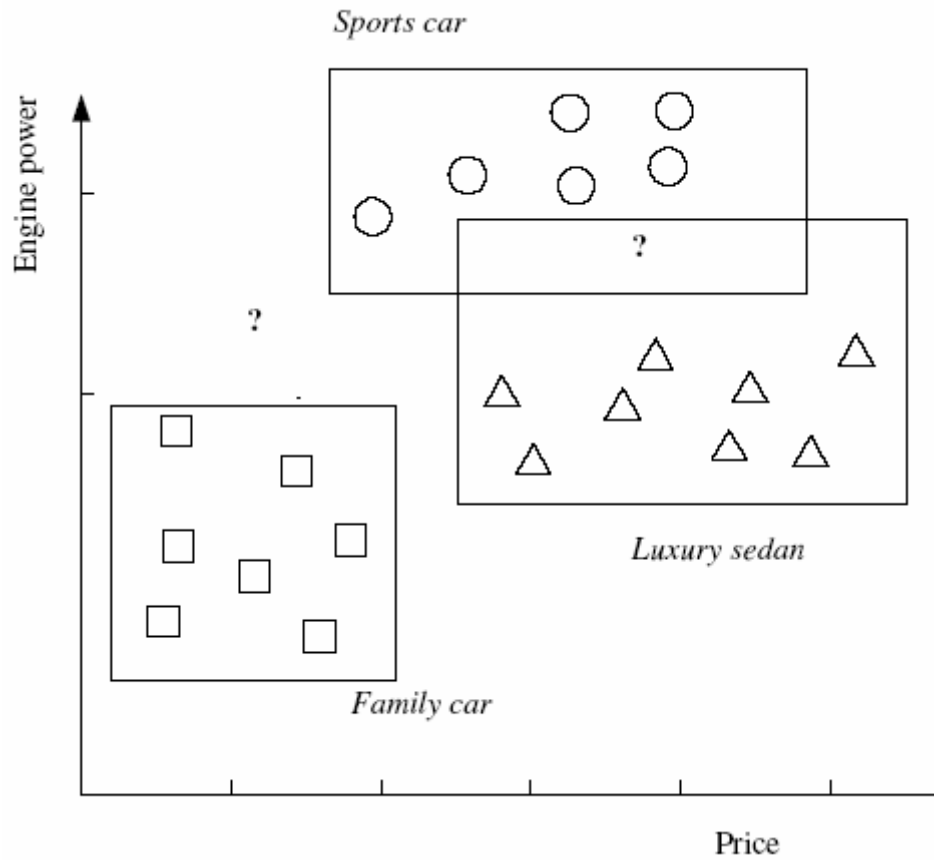
- There may not have a simple boundary between the positive and negative instances
- Zero (training) misclassification error may not be possible

Something about Simple Models

- Easier to classify a new instance
- Easier to explain
- Fewer parameters, means it is easier to train. The sample complexity is lower.
- Lower variance. A small change in the training samples will not result in a wildly different hypothesis
- High bias. A simple model makes strong assumptions about the domain; great if we're right, a disaster if we are wrong.

  optimality?: *min* (variance + bias)
- May have better generalization performance, especially if there is noise.
- Occam's razor: simpler explanations are more plausible

Learning Multiple Classes



- K-class classification

$\Rightarrow$ K two-class problems

(one against all)

$\Rightarrow$ could introduce *doubt*

$\Rightarrow$ could have unbalance data

Regression

- Supervised learning where the output is not a classification (e.g. 0/1, true/false, yes/no), but the output is a real number.
  - $\mathcal{X} = \{\mathbf{x}^t, y^t\}_{t=1}^{N}, y^t \in \mathbf{R}$

Regression

- Suppose that the true function is
  $y^t = f(\mathbf{x}^t) + \varepsilon$ where ε is random noise
- Suppose that we learn $g(x)$ as our model. The empirical error on the training set is

$$\frac{1}{N}\sum_{t=1}^{N} L(y^t, g(\mathbf{x}^t))$$

⇒ Because y^t and $g(\mathbf{x}^t)$ are numeric, it makes sense for L to be the distance between them.

⇒ Common distance measures:

- mean squared error

$$\frac{1}{N}\sum_{t=1}^{N} (y^t - g(\mathbf{x}^t))^2$$

- absolute value of difference
- etc.

Example: Linear Regression

- Assume $g(x)$ is linear

$$g(x) = w_1 x_1 + \cdots + w_d x_d + w_0 = \sum_{i=1}^{d} w_i x_i + w_0$$

and we want to minimize the mean squared error

$$\frac{1}{N} \sum_{t=1}^{N} (y^t - g(x^t))^2$$

- We can solve this for the w_i that minimizes the error

Model Selection

- Learning problem is ill-posed
- Need **inductive bias**
  - assuming a hypothesis class
  - example: sports car problem, assuming most specific rectangle
  - but different hypothesis classes will have different capacities
    - higher capacity, better able to fit the data
    - but goal is not to fit the data, it's to generalize
      - how do we measure? **cross-validation**: Split data into training and validation set; use training set to find hypothesis and validation set to test generalization. With enough data, the hypothesis that is most accurate on validation set is the best.
      - choosing the right bias: **model selection**

Underfitting and Overfitting

- Matching the complexity of the hypothesis with the complexity of the target function
  - if the hypothesis is less complex than the function, we have **underfitting**. In this case, if we increase the complexity of the model, we will reduce both training error and validation error.
  - if the hypothesis is too complex, we may have **overfitting**. In this case, the validation error may go up even the training error goes down. For example, we fit the noise, rather than the target function.

Tradeoffs

- (Dietterich 2003)
- complexity/capacity of the hypothesis
- amount of training data
- generalization error on new examples

Take Home Remarks

- What is the hardest part of machine learning?
  - selecting attributes (representation)
  - deciding the hypothesis (assumption) space: big one or small one, that's the question!
- Training is relatively easy
  - DT, NN, SVM, (KNN), …
- The usual way of learning in real life

$\Rightarrow$ not supervised, not unsupervised, but semi-supervised, even with some taste of reinforcement learning

Take Home Remarks

- Learning == Search in Hypothesis Space
- Inductive Learning Hypothesis: Generalization is possible.
- If a machine performs well on most training data AND it is not too complex, it will probably do well on similar test data.
- Amazing fact: in many cases this can actually be proven. In other words, if our hypothesis space is not too complicated/flexible (has a low capacity in some formal sense), and if our training set is large enough then we can bound the probability of performing much worse on test data than on training data.
- The above statement is carefully formalized in 40 years of research in the area of learning theory.