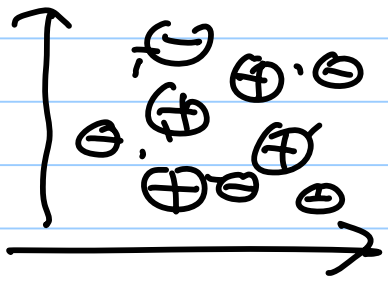
(1) # Supervised Learning. (cont)

Note Title 3/26/2008

Chp2 (cont)

Noise – any anomaly in the data.
Difficult/impossible to get zero error.

(•) Imprecision in recording input attributes.
Data points may be in incorrect positions.
(•) Errors in labeling the data points
– teacher noise.
(•) There may be additional hidden attributes that affect the label of an instance



no simple rectangle is able to seperate the +ve examples ⊕ from the -ve examples ⊖.

A more complex classifier may be needed. But there are tradeoffs.

The rectangle classifier:
(1) is simple to use.
(2) is simple to train/learn with few parameters
(3) is simple to explain.

Occam's Razor – simpler explanations are more plausible.

(2) Learning Multiple Classes

In general, we have K classes

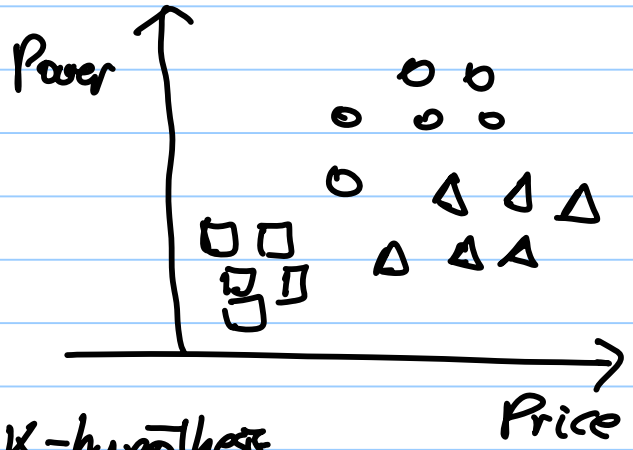
$C_i : i = 1, 2, \ldots K.$

Training set $X = \{\underline{x}^t, \underline{r}^t\}_{t=1}^N$

$$r_i^t = \begin{cases} 1, & \text{if } x^t \in C_i \\ 0, & \text{if } x^t \in C_j : j \neq i. \end{cases}$$

$\underline{r} = (r_1, r_2 .., r_K)$

□ - Family Car.
○ - Sports Car.
△ - Luxury Sedan.



We need to learn a K-hypothesis

$$h_i(x^t) = \begin{cases} 1 & \text{if } x^t \in C_i \\ 0 & \text{if } x^t \in C_j, j \neq i. \end{cases}$$

For most data, there will be a unique label
For some data, there may be no clear class assignment.

(3) Regression

Output is a numeric value (i.e. not class)

Training Set $X = \{x^t, r^t\}_{t=1}^N$ $r^t \in \mathcal{R}$

called interpolation if no noise.

Find a function $f(x)$ s.t. $r^t = f(x^t)$

Polynomial interpolation, given N points, we find the $(N-1)$st degree polynomial.

Regression, noise is added.

$r^t = f(x^t) + \varepsilon^t$ ε^t - random noise.

Noise - unknown hidden variables.

$r^t = f^*(x^t, z^t)$ z^t - hidden variables

Want to approximate the output by a function $g(x)$

Empirical error: $E(g|X) = \frac{1}{N} \sum_{t=1}^{N} [r^t - g(x^t)]^2$

(4) Regression (cont)

Goal: find $g(\cdot)$ that minimizes the empirical error.

Assume a hypothesis class for $g(\cdot)$

E.G. $g(\cdot)$ is linear

$$g(\underline{x}) = w_1 x_1 + \dots + w_d x_d + w_0 = \sum_{j=1}^{d} w_j x_j + w_0$$

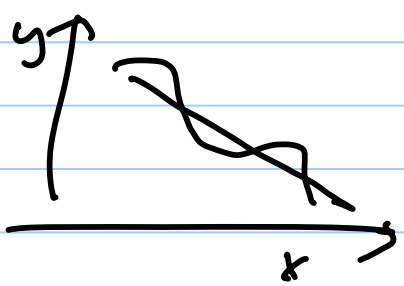
Back to simple case — single attribute x_1

$$E(w_1, w_0 | X) = \sum_{t=1}^{N} \left[r^t - (w_1 x_1^t + w_0) \right]^2$$

Minimize: $\frac{\partial E}{\partial w_1} = 0$ & $\frac{\partial E}{\partial w_0} = 0$

solution
$$\begin{cases} \hat{w}_1 = \dfrac{\sum_t x^t r^t - \bar{x}\bar{r}N}{\sum_t (x^t)^2 - N\bar{x}^2} \\ \hat{w}_0 = \bar{r} - w_1 \bar{x} \end{cases}$$

where $\bar{x} = \sum_t x^t / N$, $\bar{r} = \sum_t r^t / N$.



A "richer" model can be used.

→ e.g. $g(x) = w_2 x^2 + w_1 x + w_0$

Too high an order follows the data too closely.

(5) # Model Selection and Generalization

Consider learning a Boolean function

If d inputs, 2^d examples (at most)

Each example can be labeled 0 or 1.

Hence 2^{2^d} possible Boolean functions of d variables

Each training example removes half the hypotheses (see table)

Learning as a way to remove hypotheses inconsistent with the data

But we need to see 2^d examples to learn!

Ill-posed Data is not sufficient to determine the hypothesis (unless we have all the data)

Learning is ill-posed, we cannot learn unless we have an inductive bias.

→ assumptions about the set of hypotheses

But each hypothesis class has finite capacity and can only learn some functions.

16) Model Selection (cont)

Learning needs an inductive bias.

Model selection: How to choose the right bias?

Want the model to be able to generalize – predict new data – even more than fitting the training dataset.

Best generalization requires matching the complexity of the hypothesis with the complexity of the function underlying the data.

Underfitting – if hypothesis too simple.
Overfitting – if hypothesis too complex.

Triple Trade Off between three factors:

- The complexity of the hypothesis we fit to data, the capacity of the hypothesis class
- The amount of training data
- The generalization error on new examples

As training data increases, generalization error decreases.
As complexity of model increases, generalization decreases then increases.

(7) Model Selection (cont)²

To measure generalization ability, we need new data — validation set.
Or divide dataset into two parts
~ training data and validation set.

Cross-validation — pick hypothesis that does best on validation set

Use — training set to create hypotheses
validation set to do model selection
test set to evaluate performance.

(2) Dimensions of Supervised Machine Learning

Recapitulate & Generalize.

Sample $X = \{x^t, r^t\}_{t=1}^N$

Sample is I.I.D. Independent & Identically Distributed
All instances drawn from the same, but unknown, distribution $p(x, r)$.

Need to choose: — parameter.

(i) the model $g(x|\theta)$
defines the hypothesis class

(ii) loss function — error function.
$E(\theta|X) = \sum_t L(r^t, g(x^t|\theta))$

(iii) optimization procedure
to find $\theta^* = \arg\min_\theta E(\theta|X)$

Following conditions required.

- Hypothesis class $g(.)$ must be big enough.
- Enough training data to find best hypothesis
- Need good optimization method.