(1)

Supervised Learning

Note Title 3/26/2008

Alpaydin Chp 2.

Simplest Case – learning a class.
learning multiple classes
regression.

Learning a class from examples

class C "family car"
positive examples & negative examples

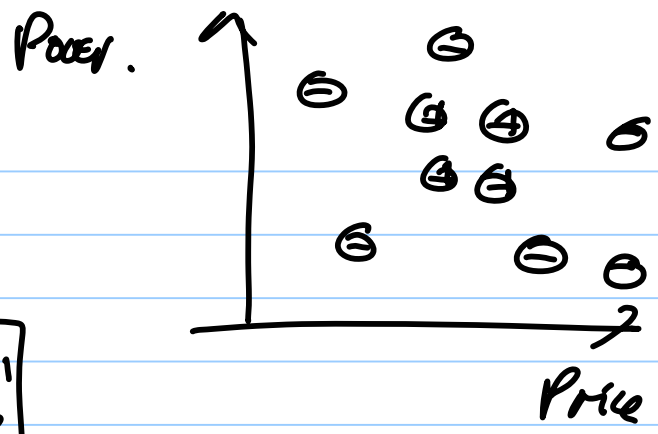
Find a description that is shared by all positive examples and no negative examples.

Predict – given a new car, is it a family car or not?

Knowledge Extraction – what do people expect from a family car?

Input Representation
– after talking to experts, decide on input representation – price, engine power.

(2)



⊕ positive example
⊖ negative example

x_1 – price
x_2 – power

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

label (class)

$$r = \begin{cases} 1, & \text{if } \underline{x} \text{ is positive example} \\ 0, & \text{if } \underline{x} \text{ is negative example} \end{cases}$$

Each car is represented by an ordered pair $(\underline{x}, r)$

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

t indexes examples.

Plot training set in 2D.

Select form of classifier – after discussion with expert

(*) $(p_1 \le \text{price} \le p_2)$ AND $(e_1 \le \text{power} \le e_2)$

for suitable values of p_1, p_2, e_1, e_2.

Assume the decision boundary is a rectangle.

Equation (*) determines the hypothesis class $\mathcal{H}$ from which to approximate the class C.

(i.e. C is the "true" class – which we don't know)

(3)

Learning is reduced to finding the best values of (p_1, p_2, e_1, e_2).

Find $h \in \mathcal{H}$

with $h(x) = \begin{cases} 1, & \text{if } x \text{ classifies } x \text{ as pos} \\ 0, & \text{if } x \text{ classifies } x \text{ as neg.} \end{cases}$

Empirical Error

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

Indicator Function.

$I(a \neq b) = 1$, if $a \neq b$, $I(a \neq b) = 0$, if $a = b$

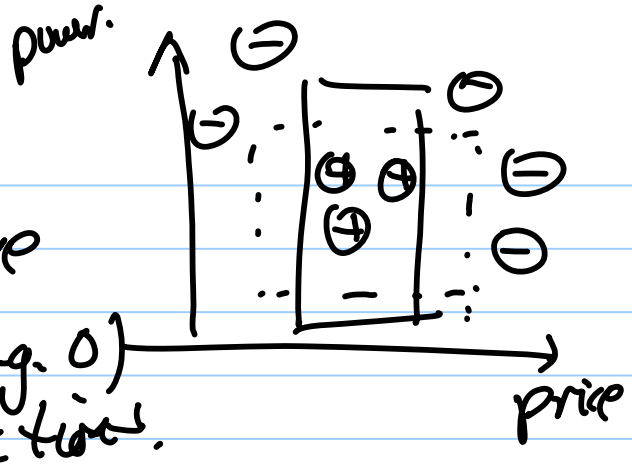
Hypothesis class — set of all possible rectangles $(p_1^h, p_2^h, e_1^h, e_2^h)$ — one hypothesis.

Task: minimize the empirical error by searching over $(p_1^h, p_2^h, e_1^h, e_2^h)$

Note: if x_1 & x_2 are real-valued, then there are an infinite number of hypotheses that will minimize the error (not true, if x_1 & x_2 are integers)

But these different "optimal" hypotheses will make different predictions on new data.

(4)



There can be many hypotheses that have the same error on the training data (e.g. 0) but make different predictions.

We care more about predictions than the results on the training dataset.

Generalization — how well will the hypothesis correctly classify future examples which are not in the training dataset.

Most specific — one possibility is to find the most specific hypothesis S (e.g. tightest rectangle) — another is the most general hypothesis (largest rectangle)

Question: Is there a classifier $h \in \mathcal{H}$ s.t. $E(h|\mathcal{X}) = 0$? e.g. $h = C$.?

Given a hypothesis space $\mathcal{H}$, there may not exist a classifier with $E(h|\mathcal{X}) = 0$.

Does $\mathcal{H}$ have enough capacity to learn C?

(5)

Vapnik-Chervonenkis (VC) Dimension.

Dataset of N points. They can be labeled in 2^N ways (positive & negative).

We can define 2^N different learning problems by these datapoints (all possible labelings).

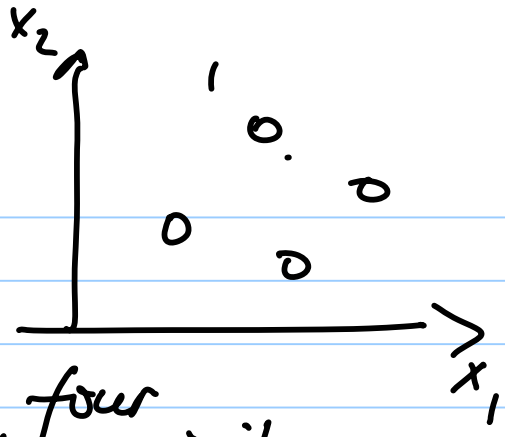
If for (all) of these problems, we can find a hypothesis $h \in \mathcal{H}$ that separates the pos from the neg (i.e. zero error), then we say that $\mathcal{H}$ shatters N points.

i.e. we can learn each problem perfectly (zero error.) by hypothesis from $\mathcal{H}$.

Max No. of points that can be shattered by $\mathcal{H}$ is the VC dimension $VC(\mathcal{H})$.

It measures the capacity of the hypothesis class $\mathcal{H}$.

(6)



An axis-aligned rectangle can shatter four points in 2-D.

$VC(\mathcal{H}) = 4$.

(It is enough that we can find four points that can be shattered – don't require it for any four points).

We cannot shatter 5 points in 2D

VC dimension – at first seems pessimistic.

→ can we only learn datasets with four points?

No, VC dimension is a measure of capacity that is independent of the prob. distribution of the examples (dataset)

Also, we don't have to worry about all possible labelings.

In fact, we can only generalize if we have more data than the VC dimension!

(7)

Probably Approximately Correct (PAC)

Using tightest rectangle S as hypothesis - how many examples do we need?

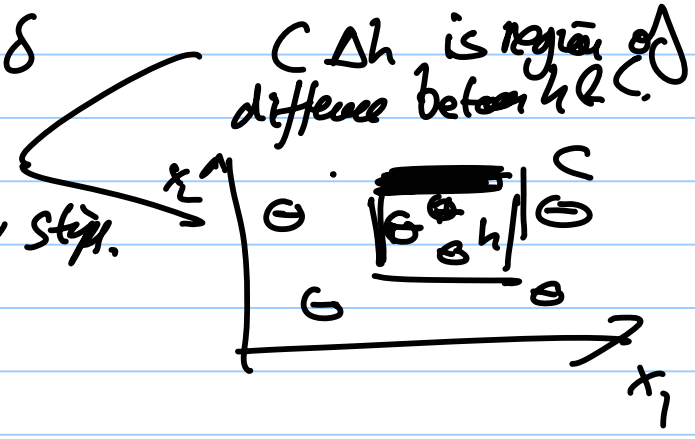
We want the hypothesis to be approximately correct - error probability is bounded by some value.

We also want our hypothesis to be be probably correct (high confidence).

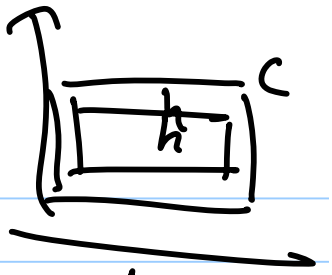
PAC given class C, examples drawn from a fixed (but unknown) distribution $p(x)$, want. no. examples N, s.t. with prob $\geq 1-\delta$, the hypothesis h has error at most ϵ (for arbitrary $\delta \leq \frac{1}{2}$ and $\epsilon > 0$)

$$P[C \Delta h \leq \epsilon] \geq 1-\delta$$

Sum of four rectangular strips.



(8) PAC (cont)



Error region between C & h=S is the sum of four rectangles.

Want prob. of +ve example falling in here (causing error) to be $\leq \epsilon$.

Want prob. $\leq \epsilon/4$ for each strip

hence prob $\leq \epsilon/4 \times 4 = \epsilon$ (double counts overlap).

Prob that random sample missing strip is $1-\epsilon/4$

prob. N samples miss is $(1-\epsilon/4)^N$.

prob N samples miss any of the four strips $\leq 4(1-\epsilon/4)^N$

want this to be at most δ

Inequality $(1-x) \leq \exp(-x)$

Choose N & δ s.t. $4e^{-\epsilon N/4} \leq \delta$

Gives $N \geq (4/\epsilon)\log(4/\delta)$.

Hence if we take at least $(4/\epsilon)\log 4/\delta$ random examples from C, and use tightest rectangle, with confidence $> 1-\delta$, given point misclassified with error $< \epsilon$.

(9) (PAC) (Cont.)

By making N large, we can decrease ϵ and δ.

$N \to \infty$, rule is almost certainly correct and almost certainly applies.

Note: Vapnik has developed a beautiful theory which generalizes these ideas.

$$N \geq f(C, \epsilon, \delta)$$

C: VC-dim, ϵ: error bound, δ: prob

Beyond scope of this course. In practice, these bounds are not very useful. But the basic concepts are.

- Capacity of Hypothesis Space C
- Error Rate ϵ
- → Confidence that rule is correct δ.