CHAPTER 2:

SUPERVISED LEARNING

Learning a Class from Examples

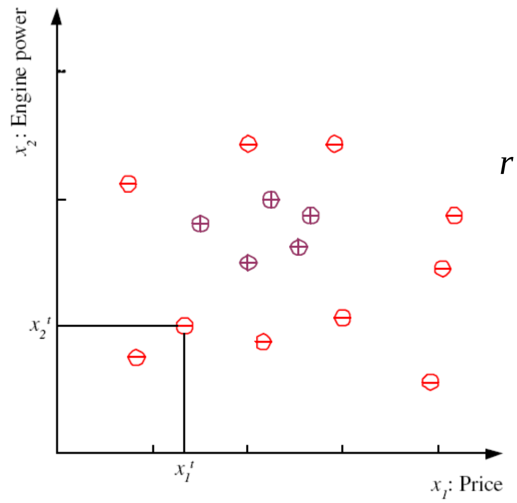
2

- Class C of a "family car"
  - Prediction: Is car *x* a family car?
  - Knowledge extraction: What do people expect from a family car?
- Output:

  Positive (+) and negative (–) examples
- Input representation:

  x_1: price, x_2 : engine power

Training set $\mathcal{X}$



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

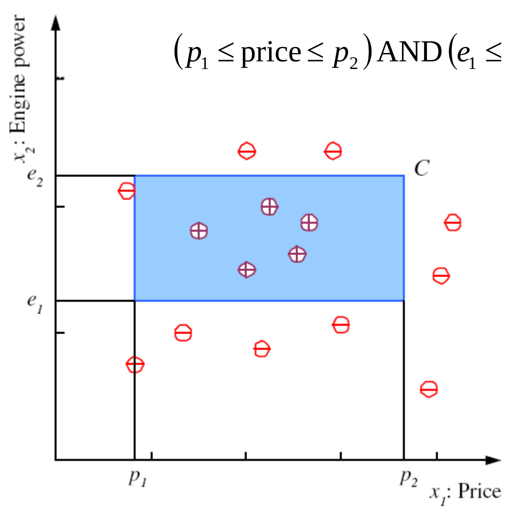
$$r = \begin{cases} 1 \text{ if } \mathbf{x} \text{ is positive} \\ 0 \text{ if } \mathbf{x} \text{ is negative} \end{cases}$$

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

Class C

4

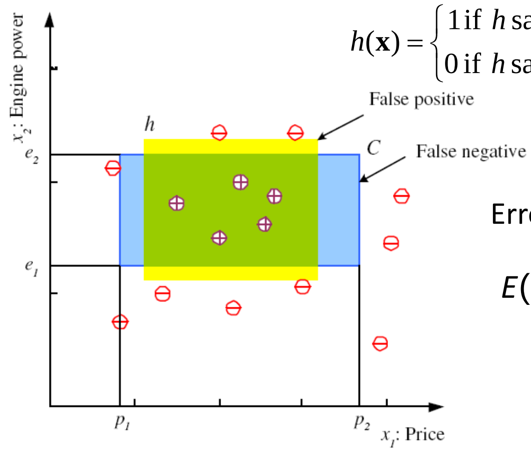
$$(p_1 \le \text{price} \le p_2) \text{ AND } (e_1 \le \text{engine power} \le e_2)$$



Hypothesis class $\mathcal{H}$

5

$$h(\mathbf{x}) = \begin{cases} 1 \text{ if } h \text{ says } \mathbf{x} \text{ is positive} \\ 0 \text{ if } h \text{ says } \mathbf{x} \text{ is negative} \end{cases}$$



Error of h on $\mathcal{H}$

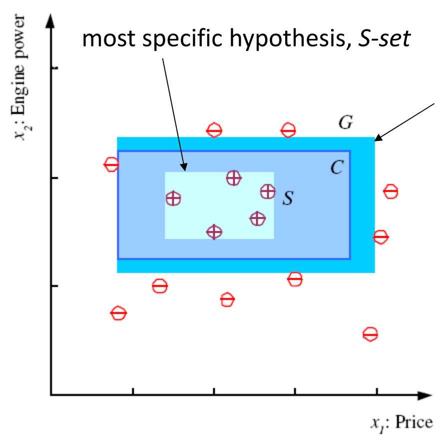
$$E(h \mid \mathcal{X}) = \sum_{t=1}^{N} 1\left(h\left(\mathbf{x}^t\right) \neq r^t\right)$$

where $I(a \neq b) = 1$ if $a \neq b$;
$= 0$ otherwise.
i.e. # of false classification

Find $h \in \mathcal{H}$, specified by $(p_1^h, p_2^h, e_1^h, e_2^h)$, to approximate C as closely as possible.

S-set, G-set, and the Version Space (AIMA 19.1)

6



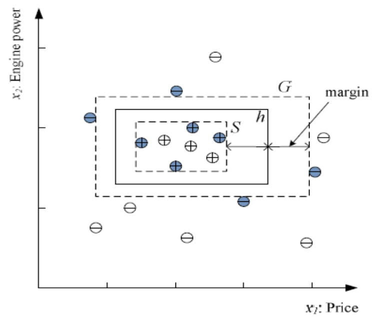
$h \in$ H, between *S-set* and *G-set* is consistent and make up the version space.

Candidate elimination: Incrementally update the S- and G-sets with the training instances one by one.

Margin

7

- Choose *h* with largest margin



- For an error function to have a minimum at *h* with the maximum margin, use an error(loss) function that checks a correct classification of an instance and its distance from the boundary.

Margin

8

- For an error function to have a minimum at *h* with the maximum margin, use an error(loss) function that checks a correct classification of an instance and its distance from the boundary. -- i.e. we need a hypothesis h that returns a value which measures the distance to the boundary and a loss function that use such *h*.
- Any instance that falls between S-set and G-set is a case of doubt – unable to label with certainty.
- Assume $\mathcal{H}$ includes C; i.e. there exists $h \in \mathcal{H}$, s.t. $E(h|\mathcal{X}) = 0$.
- Given a hypothesis class $\mathcal{H}$, we can't learn C; i.e. there exists no $h \in \mathcal{H}$ for which the error is 0.
- So, we need to make sure $\mathcal{H}$ is flexible enough, or has enough capacity to learn C.

VC Dimension

9

- *A measure of the capacity* (complexity or flexibility) of a space of functions that can be learned by a statistical classification algorithm. The capacity/flexibility of a classification model is related to how complicated it can be.
- It is defined as the *cardinality of the largest set of points that the algorithm can shatter*.
- H shatters *N* if there exists $h \in \mathcal{H}$ consistent for any of these:

  VC($\mathcal{H}$) = *N*
- the max. # of points that can be shattered by H and measures the *capacity of H.*

VC Dimension

10

- Let $\mathcal{H}$ be a set family (a set of sets) and C a set.

Their *intersection* is defined as the following set-family:

$$\mathcal{H} \cap C = \{ h \cap C \mid h \in \mathcal{H} \}$$

We say that a set C is shattered by $\mathcal{H}$ if $\mathcal{H} \cap C$ contains all the subsets of C, i.e.

$$\text{card}(\mathcal{H} \cap C) = 2^{|C|}$$

- The VC dimension of $\mathcal{H}$ is the largest integer D such that

there exists a set C with cardinality D that is shattered by H.

$$VC(\mathcal{H}) = \text{argmax}_D \ \{ |C| = D \text{ where C is shattered by H} \}$$

= the size of largest set C which is shattered by $\mathcal{H}$.

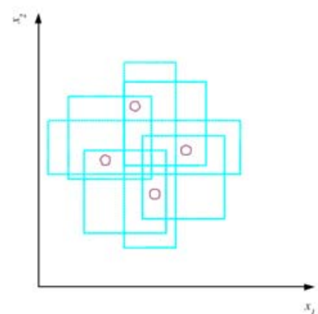
VC Dimension

11

- Vapnik-Chervonenkis Dimension
- N points can be labeled in 2^N ways as +/– ➔ 2^N different learning problems can be defined by N data points.
- $\mathcal{H}$ shatters N if there exists $h \in \mathcal{H}$ consistent for any of these: VC($\mathcal{H}$) = N

- the max. # of points that can be shattered by H and measures the *capacity of H.*

Any learning problem definable by N can be learned with no error by a hypothesis drawn from $\mathcal{H}$.

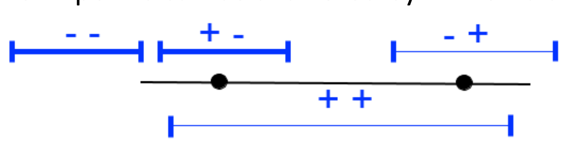


An axis-aligned rectangle shatters 4 points only !

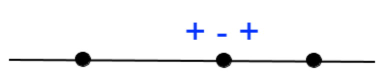
VC Dimension: Intervals of The Real Line

12

- Observations:
  - Any set of 2 points can be shattered by 4 intervals:



  - No set of 3 points can be shattered since the following dichotomy "+ - +" is not realizable (by definition of intervals):

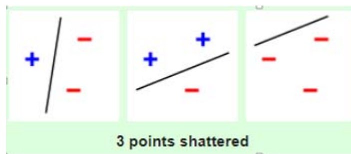


  - Thus, VC(intervals in R) = 2

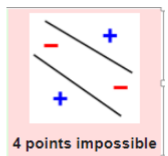
VC Dimension: Hyperplanes

13

- Observations:
  - Any three non-collinear points can be shattered:



  - Unrealizable dichotomies for four points:

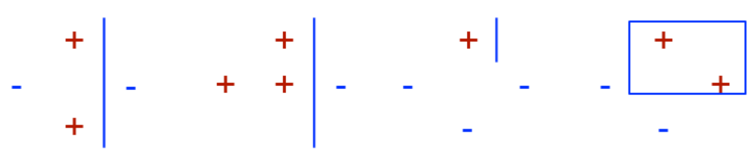


  - Thus, VC(hyperplanes in R^d) = d+1

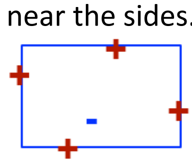
VC Dimension: Axis-Aligned Rectangles in the Plane

14

- Observations:
  - The following 4 points can be shattered:



  - No set of 5 points can be shattered: label negatively the point that is not near the sides.



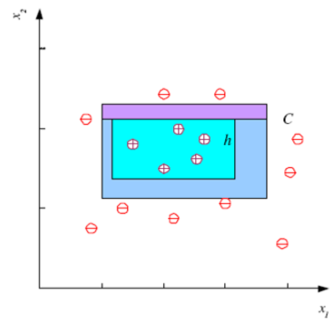
  - Thus, VC(axis-aligned rectangles) = 4

Probably Approximately Correct (PAC) Learning (AIMA 18.5, slide #34-#35)

15

- How many training examples N should we have, such that with probability at least $1 – \delta$, h has error at most ε, $\delta \le \frac{1}{2}$? (Blumer et al., '89)

  $P(C\Delta h \le \epsilon) \ge 1 - \delta$:region of difference b/t C and h.
- Each strip is at most $\varepsilon/4$
- Pr that an instance miss a strip $1– \varepsilon/4$
- Pr that N instances miss a strip $(1 – \varepsilon/4)^N$
- Pr that N instances miss 4 strips $4(1 – \varepsilon/4)^N$
- $4(1 – \varepsilon/4)^N \le \delta$ and $(1 – x) \le \exp(– x)$
- $4\exp(– \varepsilon N/4) \le \delta$ and $N \ge (4/\varepsilon)\log(4/\delta)$

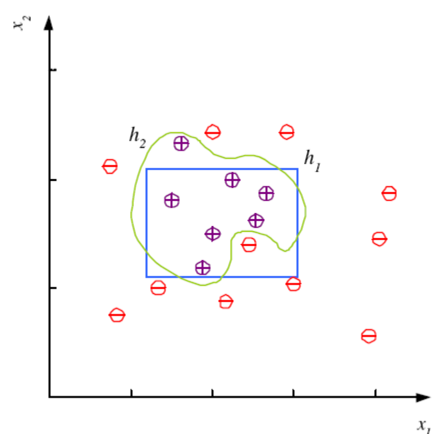


Noise and Model Complexity

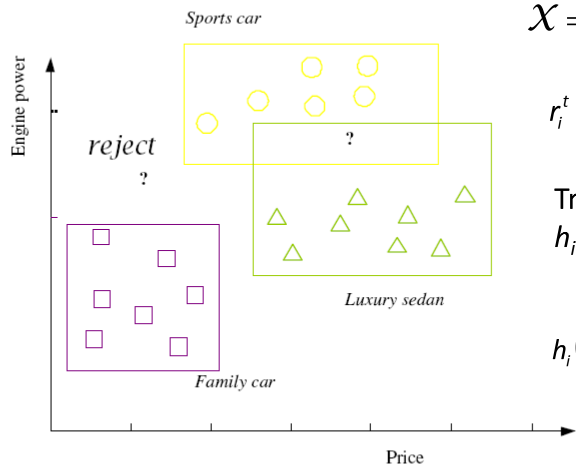
16

Use the simpler one because

- Simpler to use (lower computational complexity)
- Easier to train (lower space complexity)
- Easier to explain (more interpretable)
- Generalizes better (lower variance - Occam's razor)



Multiple Classes, C_i i=1,...,K



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

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

Train hypotheses
$h_i(\boldsymbol{x})$, i =1,...,K:

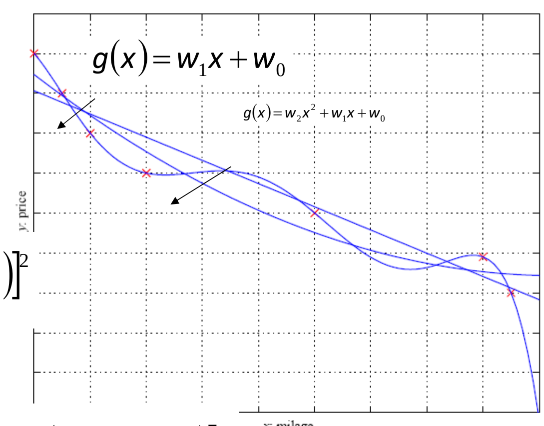
$$h_i(\mathbf{x}^t) = \begin{cases} 1 \text{ if } \mathbf{x}^t \in C_i \\ 0 \text{ if } \mathbf{x}^t \in C_j, j \neq i \end{cases}$$

Regression

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

$$r^t \in \Re$$

$$r^t = f(x^t) + \varepsilon$$



$$E(g \mid \mathcal{X}) = \frac{1}{N} \sum_{t=1}^N \left[r^t - g(x^t)\right]^2$$

i.e. Average of error

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

Model Selection & Generalization

19

- Learning is an ill-posed problem; data is not sufficient to find a unique solution
- The need for inductive bias, assumptions about $\mathcal{H}$
- Generalization: How well a model performs on new data
- Overfitting: $\mathcal{H}$ more complex than C or f
- Underfitting: $\mathcal{H}$ less complex than C or f

Cross-Validation

20

- To estimate generalization error, we need data unseen during training. We split the data as
  - Training set (50%)
  - Validation set (25%)
  - Test (publication) set (25%)
- Resampling when there is few data