Chapter 2
Supervised Learning

Learning a class from example

Class C of a “family car”

- Prediction: Is car x a family car?
- Knowledge extraction: What do people expect

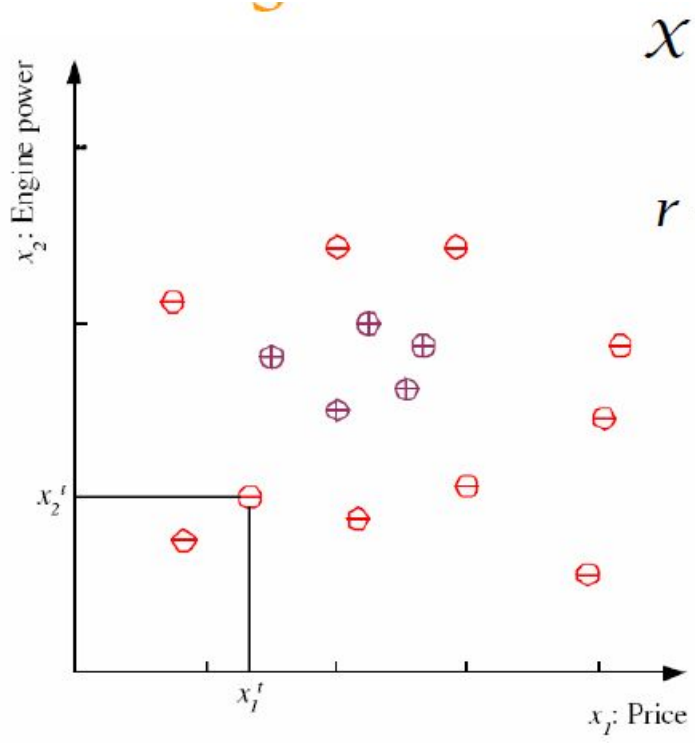
from a family car?

Output: Positive (+) and negative (–) examples

Input representation: x1: price, x2 : engine power

Seating capacity, color?

Training set X



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

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

labels

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

attributes

We want to learn the class, C, of a "family car."

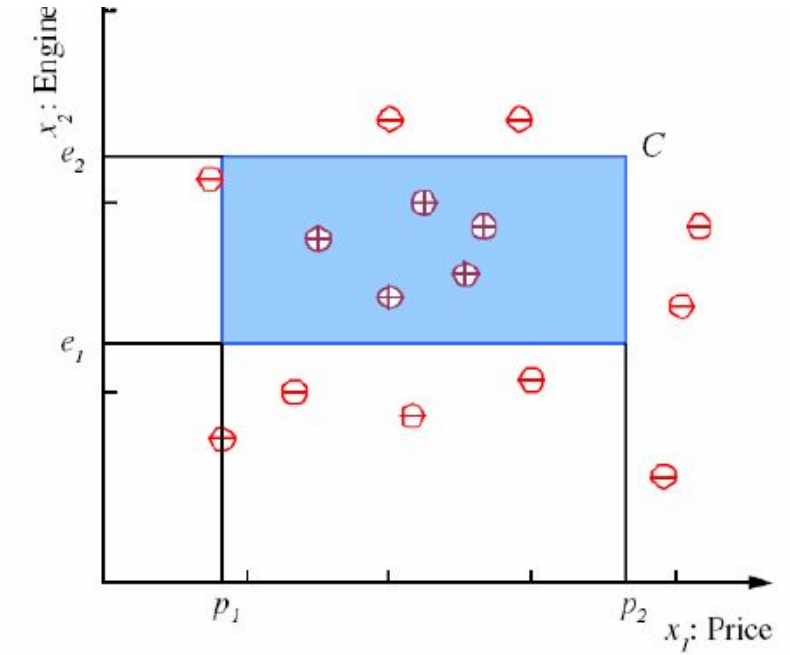
Discuss with an expert:

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

H, the hypothesis class from which we believe C is drawn, namely, the set of rectangles.

The learning algorithm then should find a particular hypothesis, h, to approximate C as closely as possible.

Which h is equal or closest to C?



Hypothesis class H

Which h is equal or closest to C?

$$h(\boldsymbol{x}) = \begin{cases} 1 \text{ if } h \text{ classifies } \boldsymbol{x} \text{ as positive} \\ 0 \text{ if } h \text{ classifies } \boldsymbol{x} \text{ as negative} \end{cases}$$

In real life we do not know C(x), so we cannot evaluate how well h(x) matches C(x).

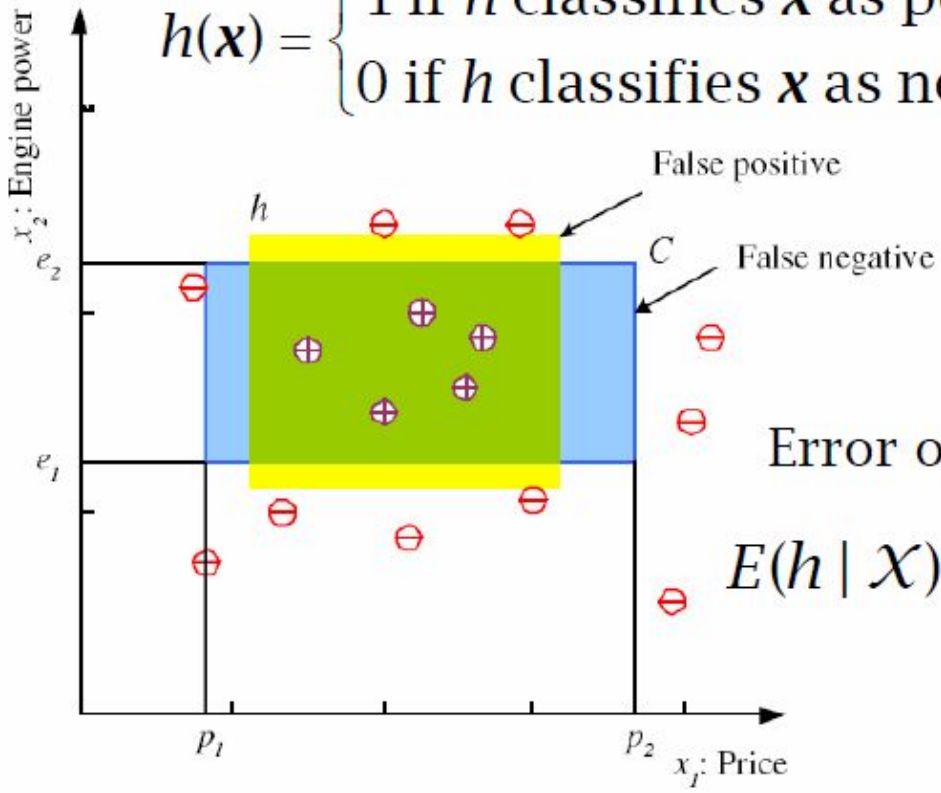
What we have is a training set X, which is a small subset of the set of all possible x

Error of h on $\mathcal{H}$

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

Hypothesis class H

$$h(\boldsymbol{x}) = \begin{cases} 1 \text{ if } h \text{ classifies } \boldsymbol{x} \text{ as positive} \\ 0 \text{ if } h \text{ classifies } \boldsymbol{x} \text{ as negative} \end{cases}$$



Error of h on $\mathcal{H}$

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

How many h(x)?

Infinite number of h(x) with zero error.

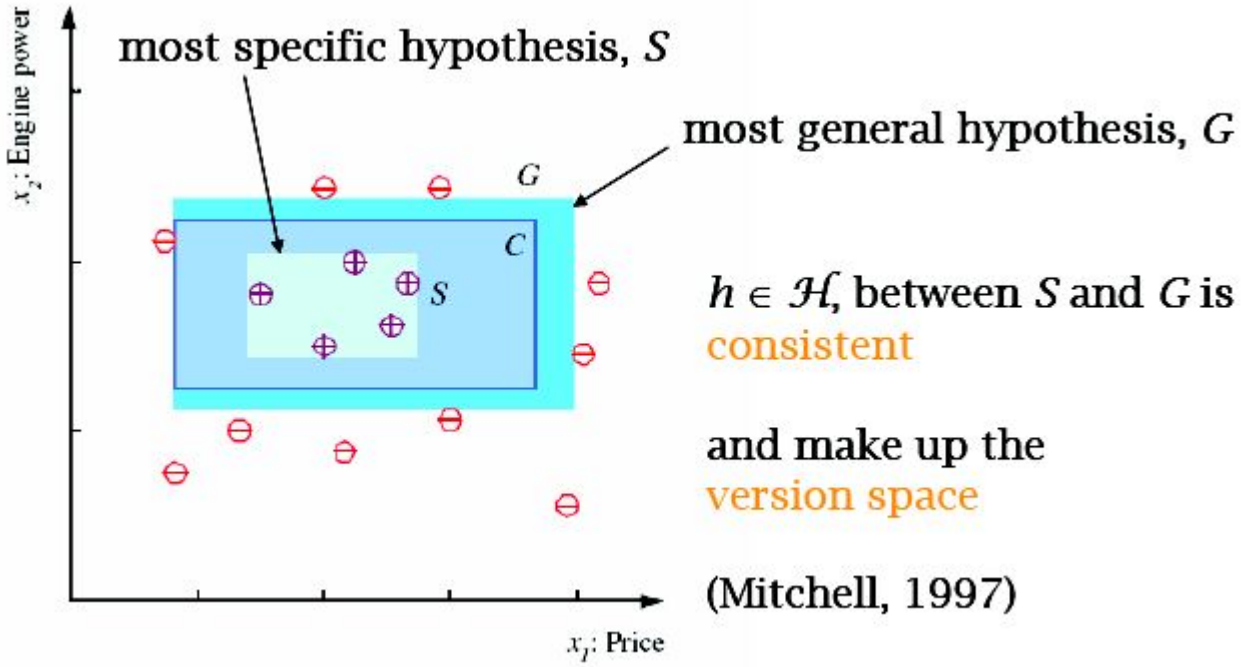
Different predictions of different candidate hypothesis.

Generalization: how well our hypothesis will correctly classify future examples that are not part of the training set.

S: most specific hypothesis

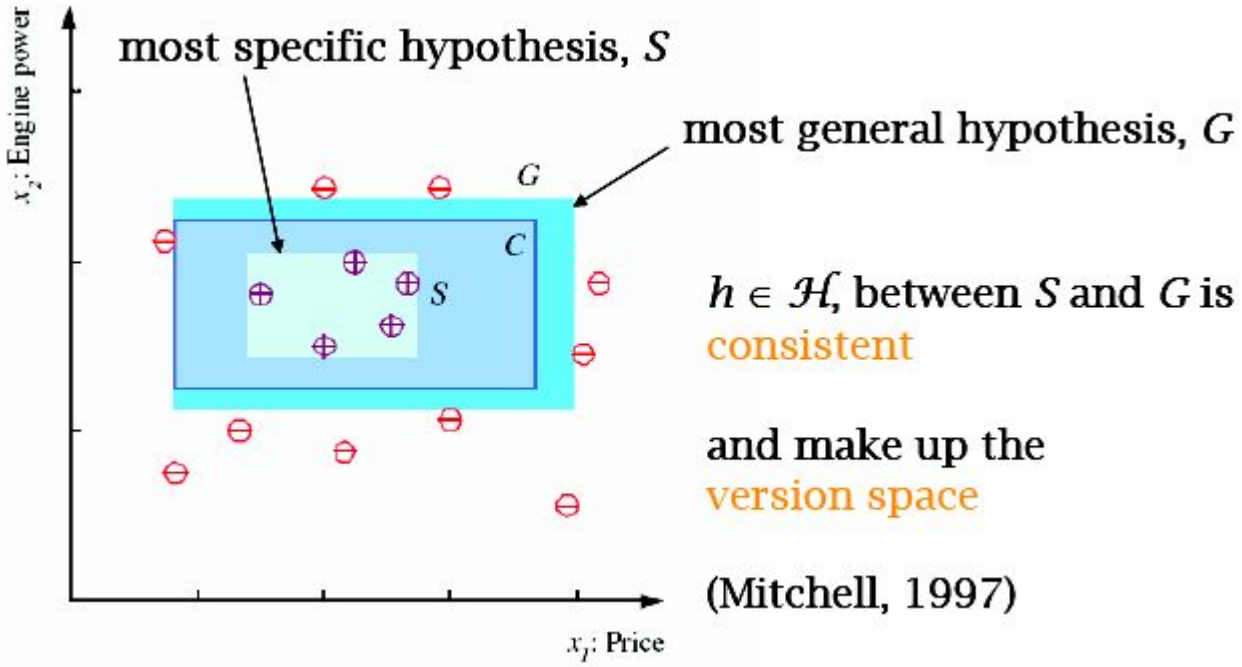
G: the most general hypothesis

S, G, and the Version Space



Any h∈H between S and G is a valid hypothesis with no error, said to be consistent with the training set, and such h make up the version space.

How to exploit S and G



Reject in case of double (uncertainty due to lack of data), defer decision to human expert

Vapnik-Chervonenkis (VC) Dimension

Make sure H is flexible enough, or has enough "capacity," to learn C.

Assume N points

How many ways to label positive or negative?

2^N learning problems can be defined.

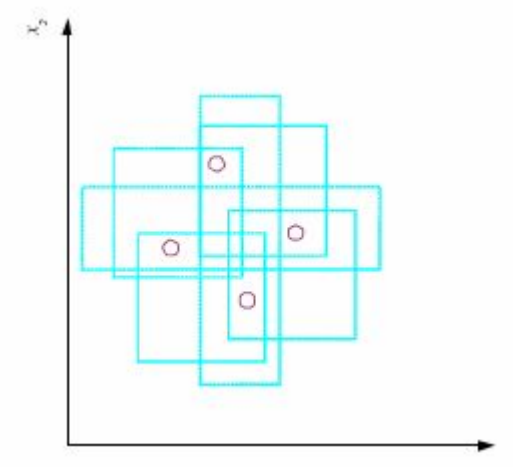
H shatters N: If any learning problem definable by N examples can be learned with no error by a hypothesis drawn from H.

The maximum number of points that can be shattered by H is called the VC dimension of H, is denoted as VC(H), and measures the capacity of the hypothesis class H.

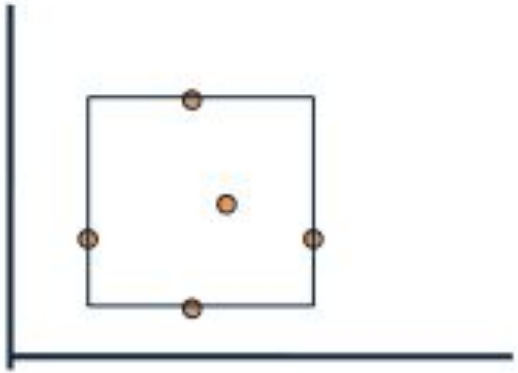
VC dimension

How many points can be shattered by an axis-aligned rectangle?

Enough that we find four points that can be shattered; it is not necessary that we be able to shatter any four points



But, no five instances can be shattered

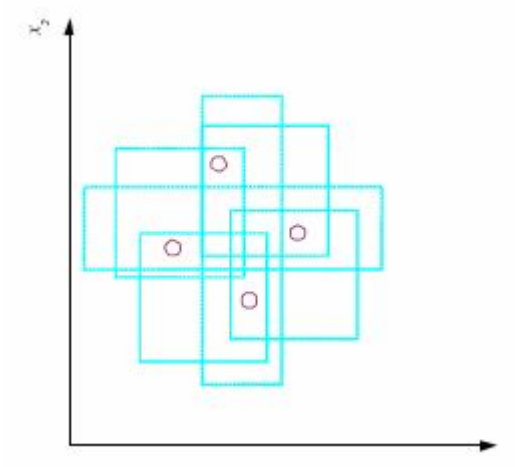


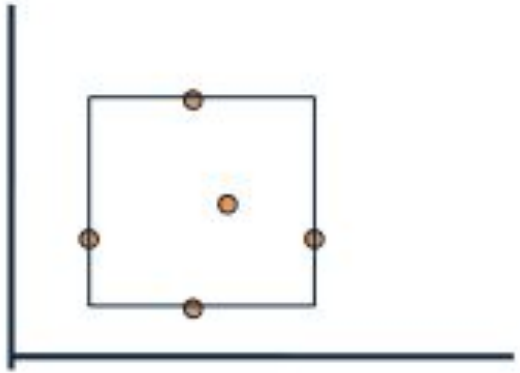
VC dimension

How many points can be shattered by an axis-aligned rectangle?

There can be at most 4 distinct extreme points (smallest or largest along some dimension) and these cannot be included (labeled +) without including the 5th point.

No five instances can be shattered





VC dimension, examples

Consider X = $\mathbb{R}$, want to learn c: X → {0,1}

What is VC dimension of

Open intervals:

H1: if x>a, then y=1 else y=0

Closed intervals:

H2: if a<x<b, then y=1 else y=0

VC dimension, examples

Let us say our hypothesis class (H) is a line instead of rectangle. How many points a line can shatter in 2 dimensional input space, i.e. what is VC(H)?

How many points a point can shatter in 1D?

Let us say our hypothesis class (H) is a convex polygon (in 2D) instead of rectangle. What is VC(H)?

VC dimension is pessimistic

- using a rectangle as our hypothesis class, we can learn only datasets containing four points and not more.
- VC dimension is independent of the probability distribution from which instances are drawn.
- In real life: world is smoothly changing, i.e. close instances have similar labels.

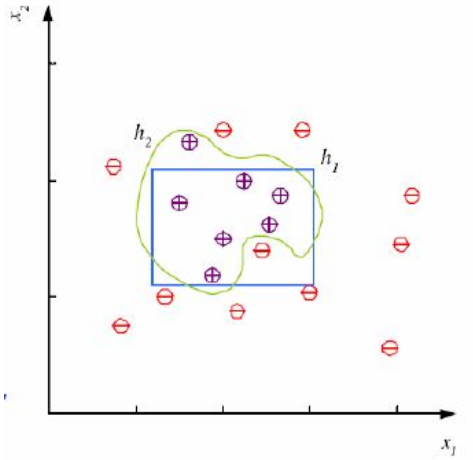
Noise and Model Complexity

With noise, might be impossible to have zero error. Several interpretations:

- Imprecision in recording the input attributes
- Error in labels
- Additional attributes that affect label. Hidden or latent. They are modelled as a random component.

With noise, there is no simple boundary.
Use simpler

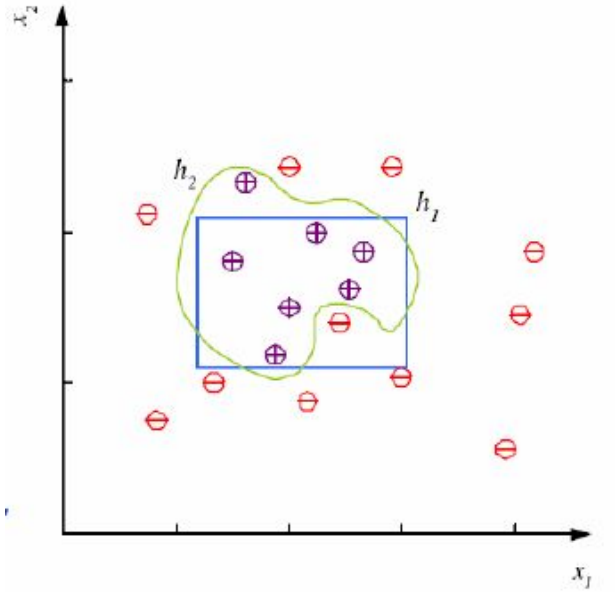
- Easier to check
- Easier to train
- Easier to explain
- Generalized better



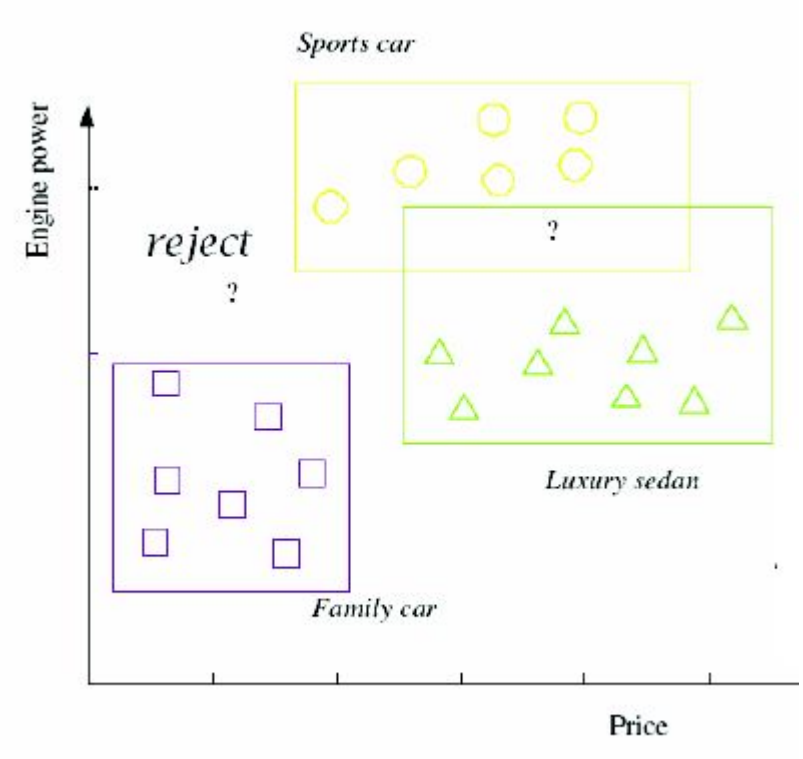
Noise and Model Complexity

Use simpler

- Easier to check
- Easier to train
  - less number of parameter
  - Small change in training instances, expect
    - Small change in simpler model
    - Large change in complex model
    - Simple model has less variance
- Easier to explain
- Generalized better
- Occam’s razor: simpler explanations are more plausible and any unnecessary complexity should be shaved off.



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



Family cars, sports cars, luxury sedan

View a K-class classification problem as K two-class problems.

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

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

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

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

Regression

The output is a numeric value, what we would like to learn is not a class, but is a continuous function.

$$\mathcal{X} = \left\{x^t, r^t\right\}_{t=1}^{N} \qquad r^t \in \Re$$

We would like to find the function f(x) that passes through these points such that we have $r^t = f(x^t)$

There is noise added to the output of the unknown function $r^t = f\left(x^t\right) + \varepsilon$

Approximate the output by our model g(x)

$E(g \mid \mathcal{X}) =$

Find g(.) that minimizes the empirical error

Regression

Find g(.) that minimizes the empirical error. Again our approach is the same; we assume a hypothesis class for g(.) with a small set of parameters. If we assume that g(x) is linear

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

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

Minimum, maximum, saddle points

What does the first derivative of f(x) at a point (p) tell about?

- If first derivative is > 0 at p, f(x) is increasing at x=p
- If first derivative is 0 at p, p is called a critical point of f(x), nothing more

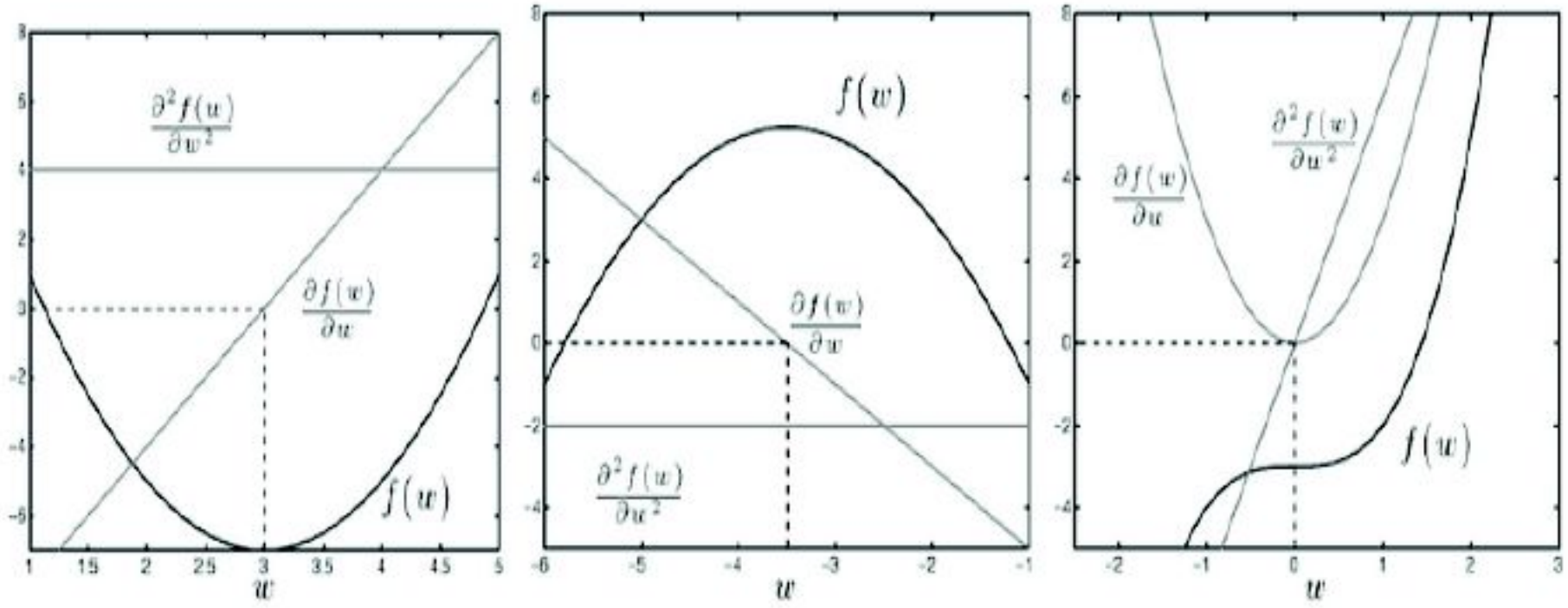
What does the first derivative of f(x) at a point (p) tell about?

- If second derivative is > 0 at p, f(x) is concave up at x=p
- If second derivative is 0 at p, then nothing more

But

- if $\frac{\mathrm{d}f}{\mathrm{d}x}(p) = 0$ and $\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}(p) > 0$
- if $\frac{\mathrm{d}f}{\mathrm{d}x}(p) = 0$ and $\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}(p) < 0$
- if $\frac{\mathrm{d}f}{\mathrm{d}x}(p) = 0$ and $\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}(p) = 0$

Minimum, maximum, saddle points



Regression

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

$$g(x) = w_1 x + w_0$$

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

$$\begin{aligned} w_1 &= \frac{\sum_t x^t r^t - \overline{xr} N}{\sum_t (x^t)^2 - N\overline{x}^2} \\ w_0 &= \overline{r} - w_1 \overline{x} \end{aligned}$$

Model selection and generalization

Learning is an ill-posed problem. d binary inputs, 2^d possible values, 2^2^d possible functions. Each distinct value removes half of the hypothesis. How many distinct values to have one hypothesis remained?

x_1	x_2	h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8	h_9	h_{10}	h_{11}	h_{12}	h_{13}	h_{14}	h_{15}	h_{16}
0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
0	1	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
1	0	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
1	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1

Model selection and generalization

Learning is an ill-posed problem. d binary inputs, 2^d possible values, 2^2^d possible functions. Each distinct value removes half of the hypothesis. How many distinct values to have one hypothesis remained?

Data is not sufficient to find a unique solution

- The need for **inductive bias**, assumptions about H

Generalization: How well a model performs on new data

- □ Overfitting: H more complex than C or f
- □ Underfitting: H less complex than C or f

Triple Trade-Off

There is a trade-off between three factors (Dietterich, 2003):

- Complexity of H, c (H),
- Training set size, N,
- Generalization error, E, on new data

- As N↑, E↓
- As c (H)↑, first E↓ and then E↑

The generalization error of an overcomplex hypothesis can be kept in check by increasing the amount of training data but only up to a point.

Cross-Validation

We can check the generalization ability of a hypothesis, namely, the quality of its inductive bias, if we have access to data outside the training set.

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%)

Dimensions of a Supervised Machine Learning Algo

We have a sample $\mathcal{X} = \{\mathbf{x}^t, r^t\}_{t=1}^N$ independent and identically distributed

Three decisions:

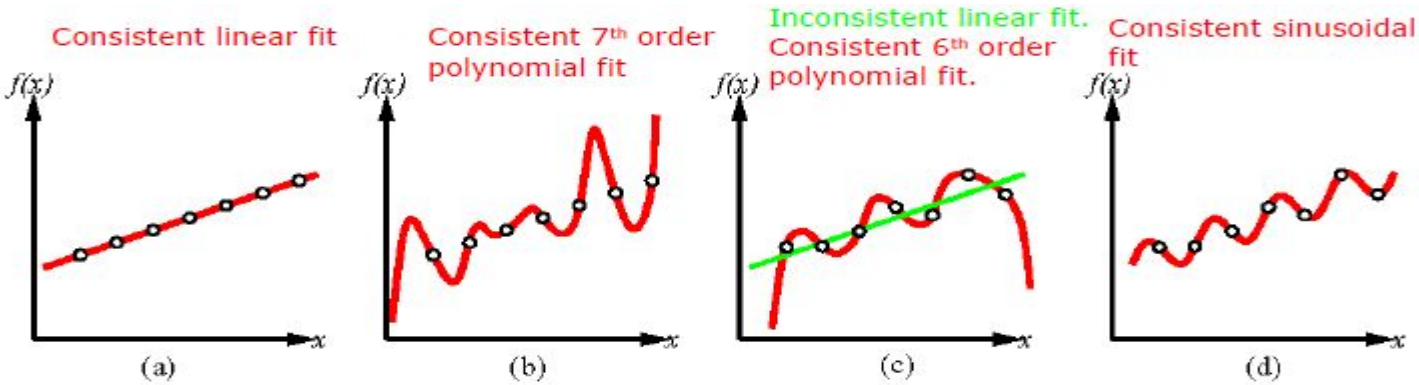
1. Model - with sufficient capacity
2. Loss function
3. Optimization procedure

$$g(\mathbf{x} \mid \theta)$$

$$E(\theta \mid \mathcal{X}) = \sum_t L(r^t, g(\mathbf{x}^t \mid \theta))$$

$$\theta^* = \arg\min_\theta E(\theta \mid \mathcal{X})$$

Model selection



Fitting a function of a single variable to some data points. f is unknown → approximate with h selected from a hypothesis space, H (e.g. the set of polynomials).

Consistent hypothesis if it agrees with all the data.
Ockham's razor: Select the simplest consistent hypothesis
Simpler hypotheses that may generalize better.

Complex model: high variance

Small change in training instances, expect large change in complex model

https://www.desmos.com/calculator