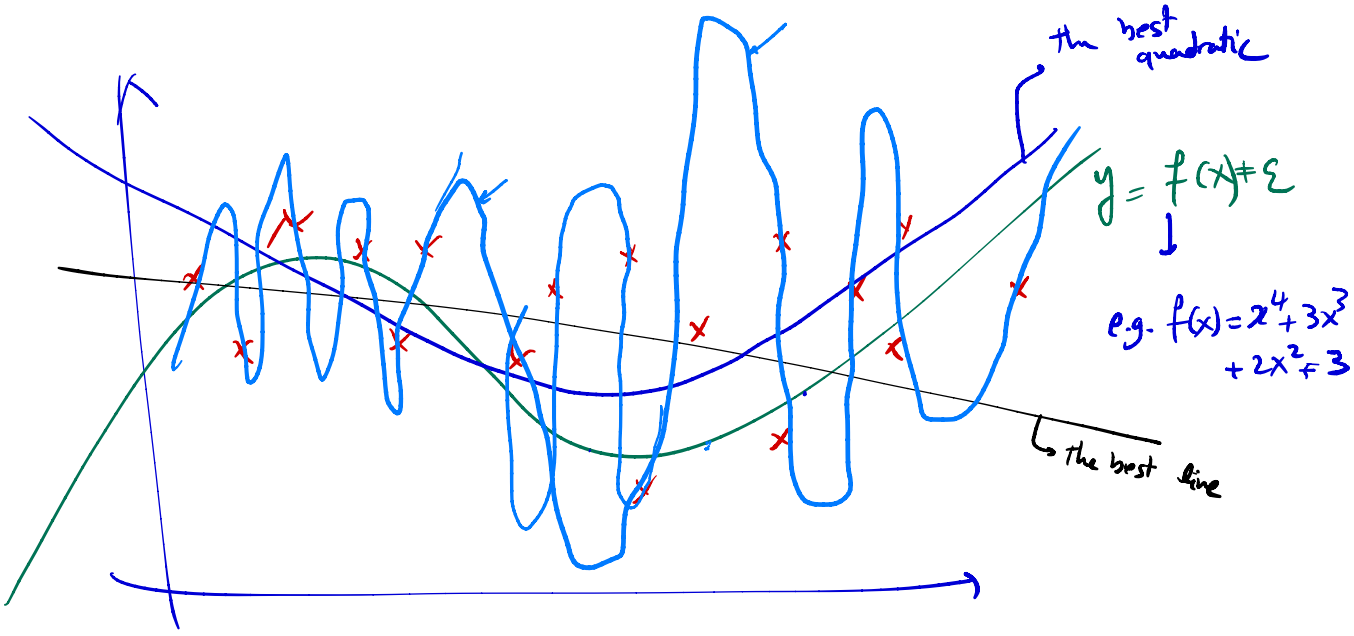
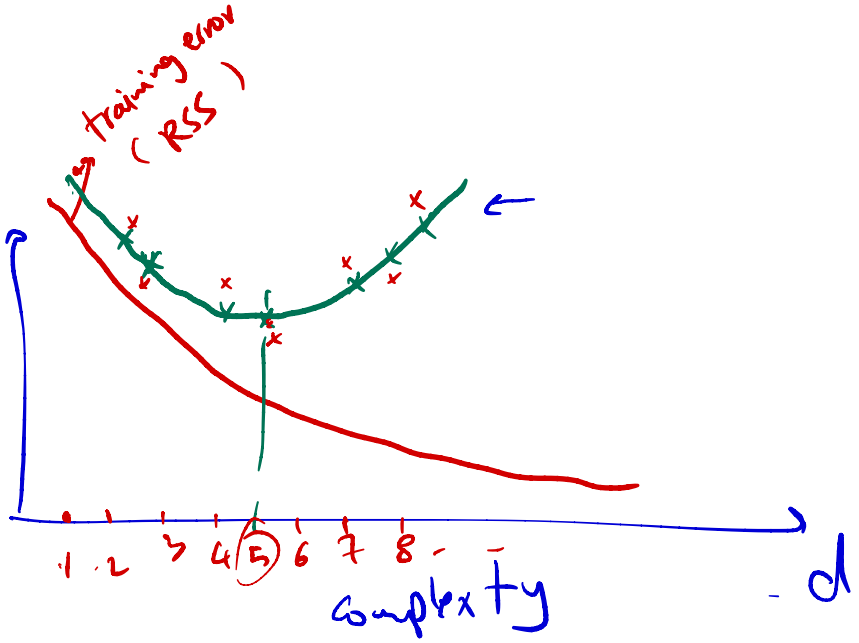
Lecture 16 :



the best fit / model with degree 100

Question : How can we estimate the the test error curve?

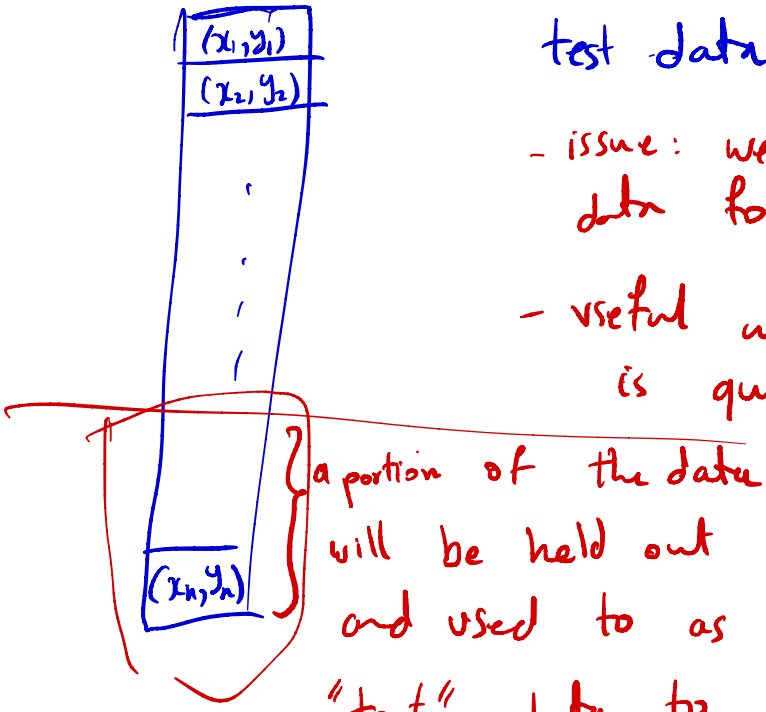


Two methods / answers:

(1) validation / hold-out data:

Randomly choose a (given) portion of the data and use it as test data. Use the rest for training.

- issue: we're not using all the available data for training
- useful when the size of the data is quite large



(2) k-fold cross-validation:

Partition/Divide the data randomly into k parts:

$$D = D_1 \cup D_2 \cup D_3 \cdots \cup D_k$$

for each $i=1,\dots,k$

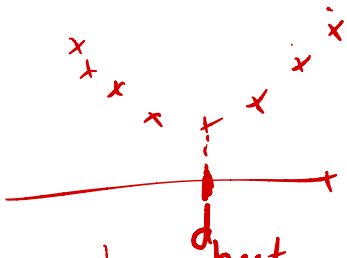
- train on $D \setminus D_i$
- test the resulting model on D_i

$\Rightarrow$ $test_i$ = test error of the trained model on D_i

$$test\text{-}error \approx \frac{1}{k}\left(test_1 + test_2 + \dots + test_k\right)$$

$\Rightarrow$ - both of the methods are used to estimate the test curve

- once the "best complexity" is found (minimum in the estimated test curve), we will use the whole data set to train a model with the "best complexity"



- Cross validation is done for each complexity to estimate the test error.

Classification:

Regression:

label $y_i \in \mathbb{R}$

classification:

label $\in$ Discrete set

$y_i \in \{0, 1\}$

$\in \{$ Spam, not-spam $\}$

$\in \{$ Sunny, Snow, rain $\}$

Example:

$\{(x_i, y_i)\}_{i=1}^{n}$

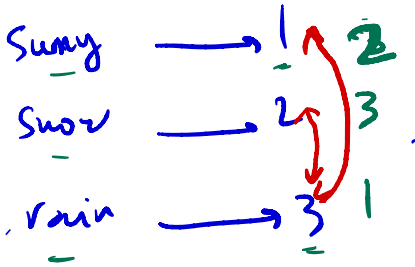
x_i = weather data for day i (clouds, ---)

$y_i = \{$ sunny, snowy, rainy $\}$

$y_i \approx f(x_i)$

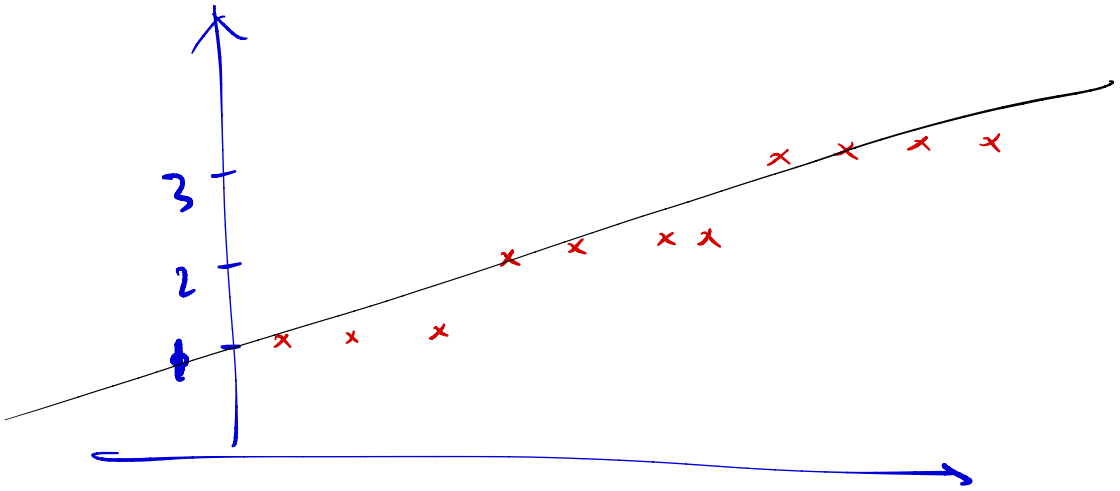
(remember, we want to be accurate at test time.

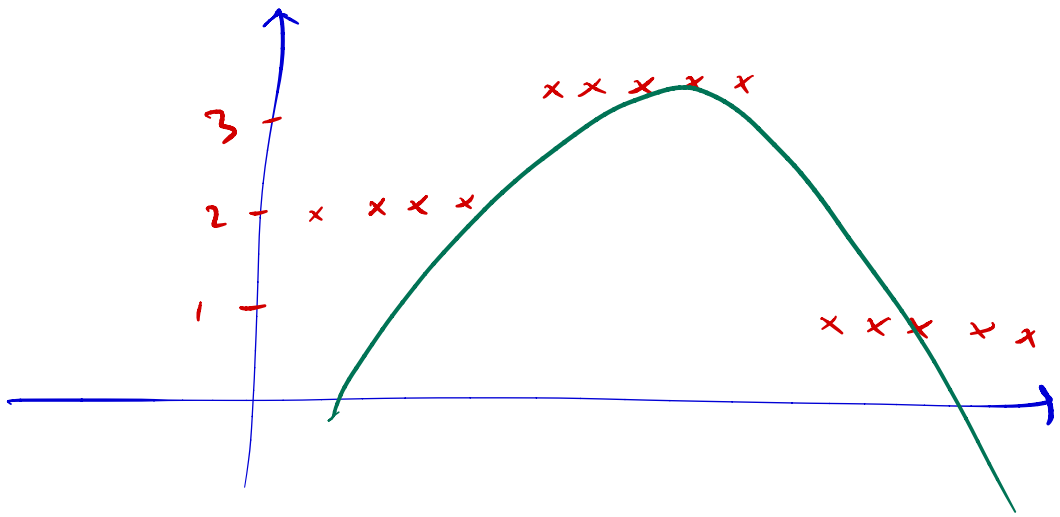
As a naive approach, let's try model/solve a classification problem using regression:



$y_i : 1, 2, 3$ → real-valued

→ let's fit a regression model





Example:

input x (input data
weather data)

$$\begin{cases} \Pr\{\text{Sunny} \mid x\} = 0.2 \\ \Pr\{\text{Snow} \mid x\} = 0.5 \\ \Pr\{\text{rain} \mid x\} = 0.3 \end{cases}$$

- what we really need to learn/estimate is the "conditional probabilities":

$$\Pr\{\text{class}_k \mid x\} \quad \text{for } k = 1, \ldots, K$$

Bayes Optimal Classifier:

Assume for the moment that all the conditional probabilities are known

$\Pr\{\text{class}_k \mid x\}$ is known for $k = 1, \ldots, K$.

How should we (optimally) predict the

class associated to input x?

Bayes optimal classifier:

$$\hat{y}_{Bayes}(x) = \text{predicted class}(x)$$

$$= \underset{k}{\text{argmax}} \; Pr\{\text{class}_k \mid x\}$$

(break ties arbitrarily)

Theorem: The Bayes optimal classifier has the smallest classification error among all the possible classifiers.

This means: for every other classifier $\hat{y}_{other}(x)$, we have:

$$Pr_{(x,y)}\{\hat{y}_{other}(x) \neq y\} \geq Pr_{(x,y)}\{\hat{y}_{Bayes}(x) \neq y\}$$

- The fundamental problem with the Bayes classifier is that it assumes the knowledge of the conditional probabilities (which not true in practice)

- All the methods for classification in ML aim at estimating the conditional probabilities.

- We'll cover several methods to estimate these conditional probabilities from data.

Logistic Regression:

For now, assume that we have two classes (the binary classification problem):

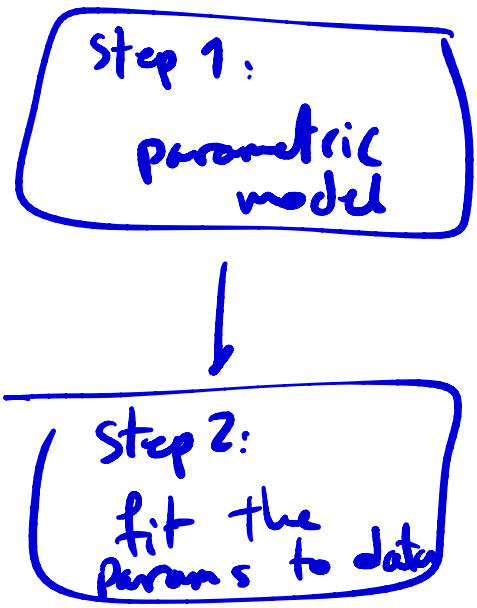
$\text{Class}_0 \longrightarrow y_i = 0$

$x_i \in \mathbb{R}$

$\text{Class}_1 \longrightarrow y_i = 1$

Goal: given $x \longrightarrow \begin{cases} \Pr\{0 \mid x\} \\ \Pr\{1 \mid x\} \end{cases}$

Data: $\{(x_i, y_i)\}_{i=1}^{n}$



Step: we're going to design a parametric model for estimation

$$\underbrace{Pr\{0|x\}}_{\in[0,1]}, \quad \underbrace{Pr\{1|x\}}_{\in[0,1]}$$

Let's begin with linear models.

What is a linear model in the binary classification setting?

$$Pr\{0|x\} = f(x; \underline{\text{parameters}})$$
$$Pr\{1|x\} = 1 - f(x; \text{parameters})$$

let's come up with simplest candidates for the parametric model.

- Perhaps the simplest model would be something like

$$\begin{cases} Pr\{0 \mid x\} = \beta_0 + \beta_1 x \\ Pr\{1 \mid x\} = 1 - (\beta_0 + \beta_1 x) \end{cases}$$

But, one can not enforce the constraint that the probabilities are between 0 and 1 using the above model (i.e. it's not possible to ensure that $\beta_0 + \beta_1 x \in [0,1]$).

Instead, we consider the following model:

$$\begin{cases} \Pr\{0|x\} = \dfrac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \\ \Pr\{1|x\} = 1-\Pr\{0|x\} = \dfrac{1}{1+e^{\beta_0+\beta_1 x}} \end{cases}$$

It's easy to see that $\dfrac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \in [0,1]$.