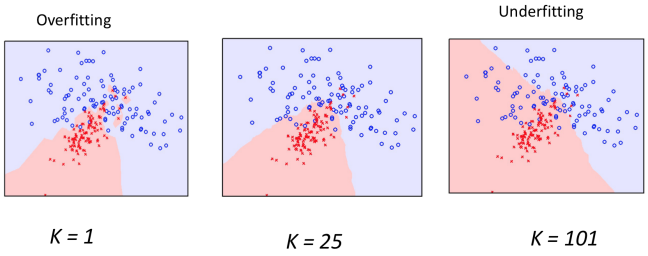
Lecture 25

§1 Overfitting and underfitting

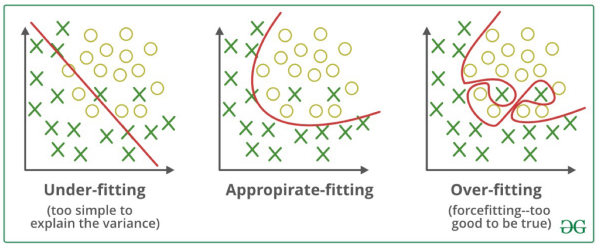
1. Overfitting (过度拟合) and underfitting (欠拟合)

- 机器学习通过向 training data 与 test data 学习，来对新的，未知的 data set 进行预测
- Overfit：机器学习了仅出现在 training set 中的 patterns，仅能对 training set 做出准确预测
- Underfit：机器无法在 training set 与 testing set 中出现的变量间找到 major patterns.

KNN



Overfitting and Underfitting



2. Example: Nonlinear regression model

- Want to fit a polynomial regression model

$$y = \theta_0 + \theta_1 x + \theta_2 x^2 + \cdots\cdots + \theta_n x^n + \varepsilon$$

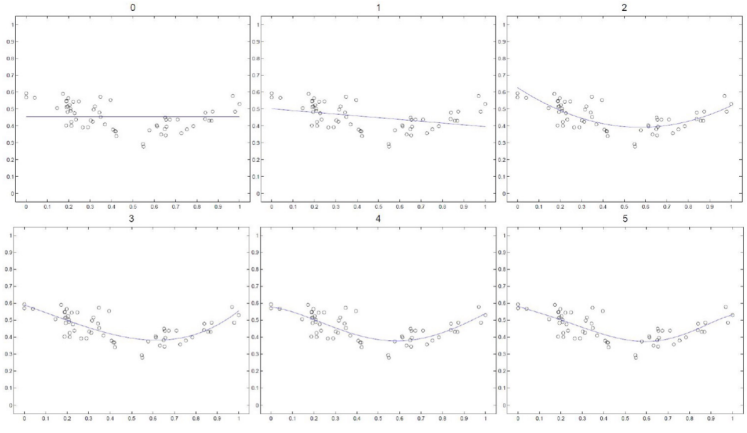
* ε 为 noise term

- 令 $x = (1, x, x^2, \cdots\cdots, x^n)^T$，$\theta = (\theta_0, \theta_1, \theta_2, \cdots\cdots, \theta_n)^T$，则

$$y = \theta^T x$$

注：增加 maximal polynomial degree 可能使曲线穿过所有的 training points，但这样会导致 overfit

Increasing the maximal degree

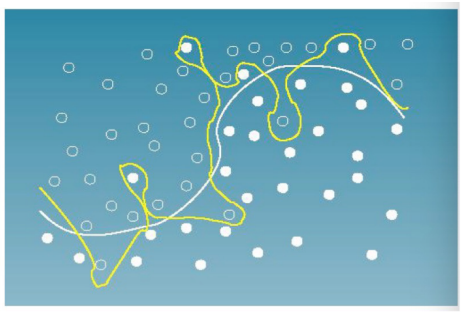


3. Example: Classification model

- Logistic regression with polynomial features

$$P(y=1 \mid x, \theta) = \frac{1}{1+\exp(-\theta^T x)}$$

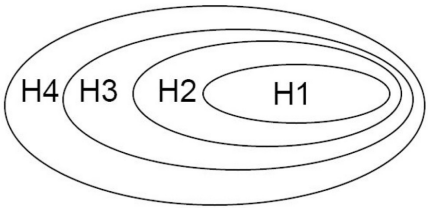
- 令 $x = (1, x, x^2, \cdots, x^n)^T$, $\theta = (\theta_0, \theta_1, \theta_2, \cdots, \theta_n)^T$



4. Model space

- Model 越复杂，model space 就越大
- e.g. 2阶的 polynomial function 包含了 1 阶的 polynomial function

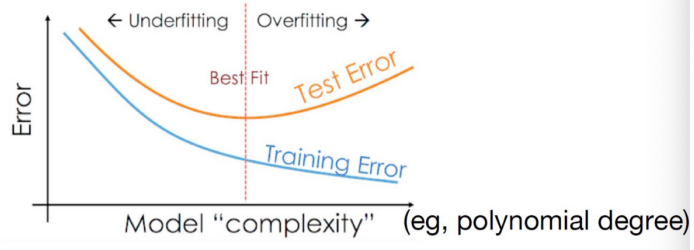
• Eg. Polynomial function of degree 1, 2, … corresponds to space H1, H2 …



§2 Select a model: K-fold validation

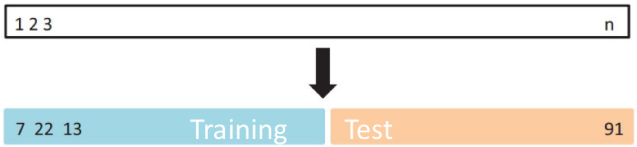
1. Intuition of model selection

• Find the right model family s.t. test error becomes minimum



2. Validation set (验证集) approach

- 将 samples 分为 training data 与 test data



- 要从一系列可能的 model 中选出一个

- 对于每一个可选的 model m
  - 用 training data 来 train model
  - 用 trained model 来预测 test data，得出 prediction error (Err_m)
- 选出 test error 最小的 model ($\min Err_m$)

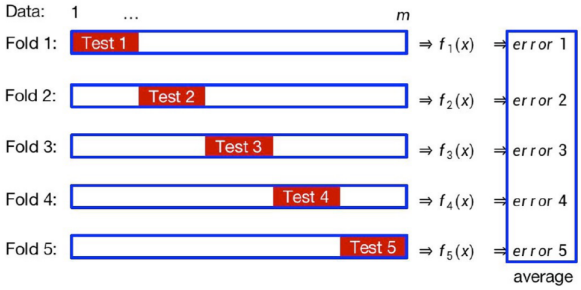
* 此方法选出的模型高度取决于 your separation of the data.
  解决方法：重复 process 多次，每次使用不同的 training and test data.

3. K-fold validation

- 将所有的 data 分为 K 个部分，每个部分均不与其他部分重复
- 建立 K 个 folds，其中第 i 个 fold 以第 i 个部分为 test data，其他部分为 training data.
- 对于每一个 model，都可以由 K 个 folds 中的 K 组 training data 训练出 K 个拟合的 function，再分别由对应的 test data 检验，可以得到 K 个 prediction error.
- 比较不同模型的 error 的平均值，选择最小的模型.

K-fold Validation

- 5-fold cross-validation (blank: training; red: test)

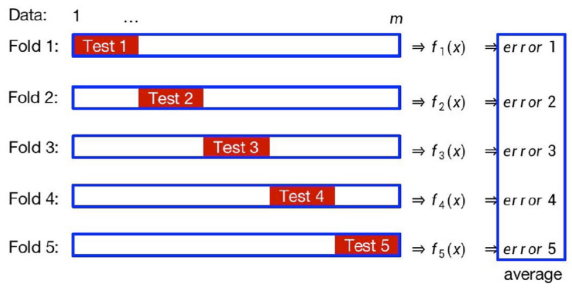


  - f_i is fitted by the training data in Fold i.
  - For each fold, use test data to test the prediction error.
  - Use the **average** error as the model's prediction error.

Use training data to train model

Model 1 $h(\theta, X)$	Model 2 $g(\gamma, X)$
$h(\theta_1, X)$	$g(\gamma_1, X)$
$h(\theta_2, X)$	$g(\gamma_2, X)$
$h(\theta_3, X)$	$g(\gamma_3, X)$
$h(\theta_4, X)$	$g(\gamma_4, X)$
$h(\theta_5, X)$	$g(\gamma_5, X)$

- 5-fold cross-validation (blank: training; red: test)



  - f_i is fitted by the training data in Fold i.
  - For each fold, use test data to test the prediction error.
  - Use the **average** error as the model's prediction error.

Error

Model 1 $h(\theta, X)$	Model 2 $g(\gamma, X)$
$Err_h(\theta_1)$	$Err_g(\gamma_1)$
$Err_h(\theta_2)$	$Err_g(\gamma_2)$
$Err_h(\theta_3)$	$Err_g(\gamma_3)$
$Err_h(\theta_4)$	$Err_g(\gamma_4)$
$Err_h(\theta_5)$	$Err_g(\gamma_5)$

Model 1 is better iff.
$\frac{1}{K}\sum_i Err_h(\theta_i) < \frac{1}{K}\sum_i Err_g(\gamma_i)$

* K 的选取：
通常情况下 K 取 10，α（每一部分的占比）$= \frac{1}{K} = 0.1$
K 太大会 time-consuming
若不考虑时间成本，K 越大越好.