



电子科技大学
University of Electronic Science and Technology of China



Criteria for model selection

Wenbao Li



Data Mining Lab, Big Data Research Center, UESTC

Email: liwenbao930116@gmail.com



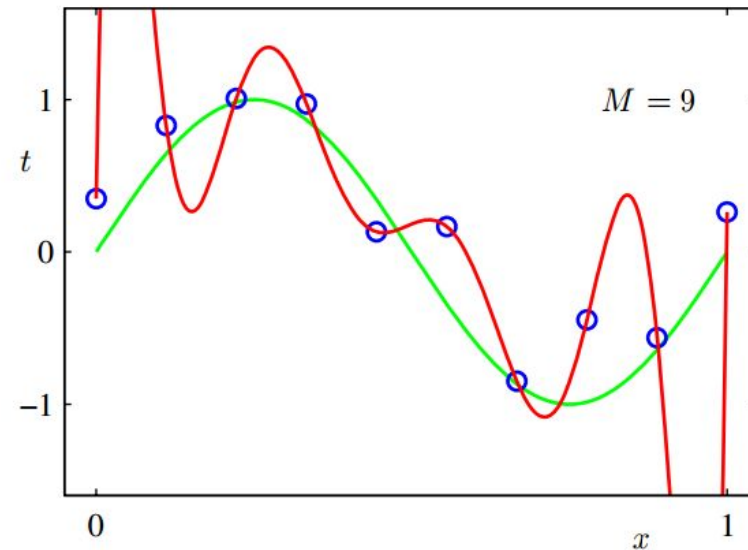
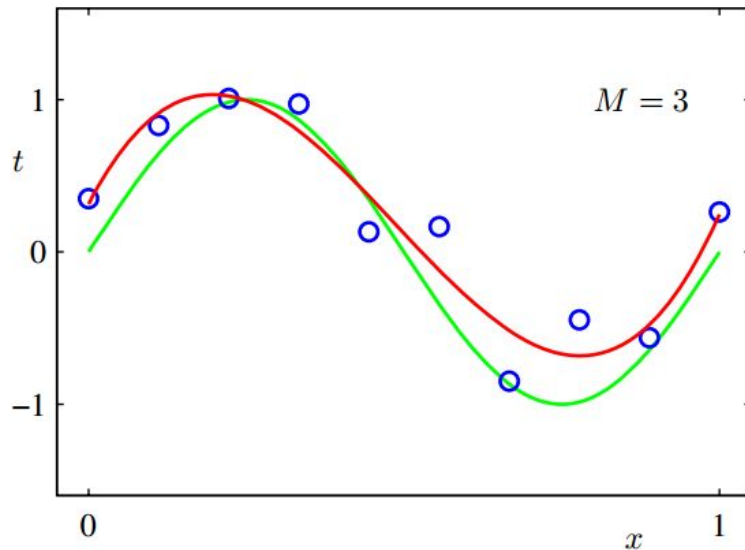
- Introduction of model selection
- The criteria for model selection
- The application



Next

An Introduction

- For example:
- Overfitting in machine learning regression problems:



$$t = \sum_{i=0}^m \theta_i x^i \quad m = ?$$

- So, how do we decide which model is to be selected?

- Selecting a **(best)** statistical model from a set of candidate models, given data.
- **goodness of fit** + **less complexity**
- **Goodness of fit** is generally determined using a [likelihood ratio approach](#), or an approximation of this, leading to a chi-squared test.
- **The complexity** is generally measured by counting the number of parameters in the model.(simple)
- And we need distinguish two goals:
 - Find the model that gives the best prediction (without assuming that any of the models are correct).(AIC, CV)
 - Assume one of the models is the true model and find the "true" model.(BIC)

- Compare the goodness of fit of two models
- based on the likelihood ratio
- Assume there are two models m_1 , the null one and alternative one m_2 . The likelihood ratio is :

$$D = -2 \ln \left(\frac{\text{likelihood}(m_1)}{\text{likelihood}(m_2)} \right) = 2 \ln(\text{likelihood}(m_2)) - 2 \ln(\text{likelihood}(m_1))$$

- According to Wilk's theorem, the p.d.f of D is χ^2 distribution with degrees of freedom equal to $d_{m_2} - d_{m_1}$
- So we can determine the *p-value* of the test.



Next

The Criteria for Model Selection

The Criteria for Model Selection



- AIC(Akaike Information Criterion)
- Bayes Factor
- BIC(Bayes Information Criterion)
- MDL(Minimum Description Length)
- Cross-validation



Next

I: Akaike Information Criterion



- AIC is founded on information theory: it offers a relative estimate of the information lost when a given model is used to represent the process that generates the data. In doing so, it deals with the trade-off between the goodness of fit of the model and the complexity of the model.

$$AIC = 2k - 2 \ln(L)$$

- L is the maximum value of the likelihood function for the model
- k is the number of estimated parameters in the model.
- the preferred model is the one with the minimum AIC value.

- It is an unbiased estimate of Kullback–Leibler divergence
- Suppose we have k models where each model is a set of densities:

$$M_j = \{p(y; \theta_j) : \theta_j \in \Theta_j\}$$

we have data Y_1, Y_2, \dots, Y_n drawn from some density f . **We do not assume that f is in any of the models.**

let $\hat{\theta}_j$ be the mle from model j . Then an estimate of p based on model j is $\hat{\Pr}_j(y) = \Pr(y; \hat{\theta}_j)$ the quality of $\hat{\Pr}_j(y)$ as an estimate of f can be measured by kl distance:

$$\begin{aligned} K(p, \hat{p}_j) &= \int p(y) \log \left(\frac{p(y)}{\hat{p}_j(y)} \right) dy \\ &= \int p(y) \log p(y) dy - \int p(y) \log \hat{p}_j(y) dy = K_j \end{aligned}$$

min max

- Intuitively, an estimate of K_j is :

$$\overline{K}_j = \frac{1}{n} \sum_{i=1}^n \log p(Y_i; \hat{\theta}_j) = \frac{\ell_j(\hat{\theta}_j)}{n}$$

- However , it is very biased because the data are being used twice: first to get the mle and second to estimate the integral. Akaike showed that the bias is approximately d_j/n where $d_j = \text{dimension}(\theta_j)$, therefore we use

$$\hat{K}_j = \frac{\ell_j(\hat{\theta}_j)}{n} - \frac{d_j}{n} = \overline{K}_j - \frac{d_j}{n}.$$

- Now ,define

$$\text{AIC}(j) = 2n\hat{K}_j = \ell_j(\hat{\theta}_j) - 2d_j.$$

- Assume there are R candidate models for selection, and their AIC values are $AIC_1, AIC_2, \dots, AIC_R$. AIC_{\min}
- Then $e^{\frac{AIC_{\min} - AIC_i}{2}}$ can be interpreted as the relative probability that the i th model minimizes the (estimated) information loss. The quantity is called relative likelihood of model i .
- Example:
 - 3 models with AIC values are 100, 102, 110. Then the relative likelihood of the second and third models are
Model 2: $\exp((100-102)/2) = 0.368$.
Model 3: $\exp((100-110)/2) = 0.007$.
 - From further consideration, remove the third model and three strategies.
 - (i) gather more data to distinguish between the first two models,
 - (ii) simply conclude data insufficient select from the first two,
 - (iii) take a weighted average of the first two models, with weights 1 and 0.368.



- If all the models in the candidate set have the same number of parameters, then using AIC might at first appear to be very similar to using the likelihood-ratio test.
- There are, however, important distinctions. In particular, the likelihood-ratio test is valid **only for nested models**, whereas AIC (and AICc) has no such restriction.

- AICc(a kind of formula)

$$AICc = AIC + \frac{2k(k+1)}{n-k-1}$$

- where n denotes the sample size and k denotes the number of parameters
- Model are univariate and linear with normal-distributed residuals.

assumption

- AICc is AIC with a **correction** for **finite sample sizes**.
- AICc is essentially AIC with a **greater penalty** for extra parameters.



Next

Bayes Factor

Bayesian Information Criterion

- Definition:
 - Posterior probability

$$\Pr(M | D) = \frac{\Pr(D | M) \Pr(M)}{\Pr(D)}$$

- So we have

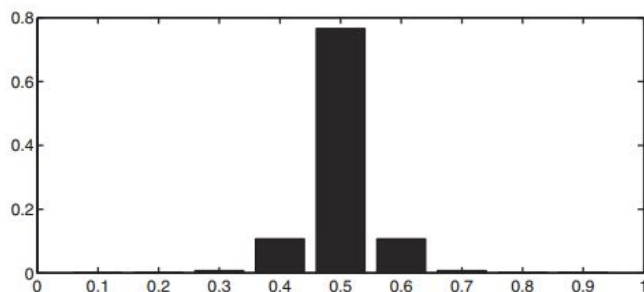
$$\frac{\Pr(M_i | D)}{\Pr(M_j | D)} = \frac{\Pr(D | M_i) \Pr(M_i)}{\Pr(D | M_j) \Pr(M_j)}$$

Posterior Odds Bayes' Factor Prior Odds

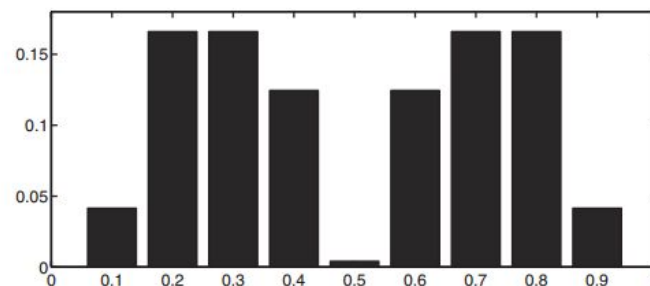
- Bayes factor

$$K = \frac{\Pr(D | M_1)}{\Pr(D | M_2)} = \frac{\int \Pr(\theta_1 | M_1) \Pr(D | \theta_1, M_1) d\theta_1}{\int \Pr(\theta_2 | M_2) \Pr(D | \theta_2, M_2) d\theta_2}$$

- Example: coin tossing
- Two competing models, one corresponding to a fair coin, and the other a biased coin.



(a)



(b)

Figure 12.1 (a) Discrete prior $p(\theta|M_{fair})$ model of a 'fair' coin. A perfectly unbiased coin has $\theta = 0.5$, which would correspond to a prior $\delta(\theta, 0.5)$ – however, we assume a more general form here to illustrate how richer prior assumptions can be used. (b) Prior $p(\theta|M_{biased})$ for a biased 'unfair' coin. In both cases we are making explicit choices here about what we consider to be 'fair' and 'unfair'.

$$p(\mathcal{D}|M) = \sum_{\theta} p(\mathcal{D}|\theta, M)p(\theta|M) = \sum_{\theta} \theta^{N_H} (1 - \theta)^{N_T} p(\theta|M) \tag{12.2.1}$$

$$= 0.1^{N_H} (1 - 0.1)^{N_T} p(\theta = 0.1|M) + \dots + 0.9^{N_H} (1 - 0.9)^{N_T} p(\theta = 0.9|M). \tag{12.2.2}$$

Example 12.1 Discrete parameter space

5 heads and 2 tails Using $N_H = 5$, $N_T = 2$ in Equation (12.2.2) we obtain $p(\mathcal{D}|M_{fair}) = 0.00786$ and $p(\mathcal{D}|M_{biased}) = 0.0072$. The posterior odds is

$$\frac{p(M_{fair}|\mathcal{D})}{p(M_{biased}|\mathcal{D})} = 1.09 \quad (12.2.3)$$

indicating that there is little to choose between the two models.

50 heads and 20 tails For this case, repeating the above calculation, we obtain $p(\mathcal{D}|M_{fair}) = 1.5 \times 10^{-20}$ and $p(\mathcal{D}|M_{biased}) = 1.4 \times 10^{-19}$. The posterior odds is

$$\frac{p(M_{fair}|\mathcal{D})}{p(M_{biased}|\mathcal{D})} = 0.109 \quad (12.2.4)$$

indicating that we have around 10 times the belief in the biased model as opposed to the fair model.

- For a model with parameter vector θ , $\dim(\theta)=K$, we have data D , then the model likelihood is

$$\Pr(D | M) = \int \Pr(D | \theta, M) \Pr(\theta | M) d\theta$$

- For a generic expression

$$\Pr(D | \theta, M) \Pr(\theta | M) = \exp(-f(\theta))$$

- Mostly, it's **difficult to evaluate the integral for large K** unless f is of a particular simple form.

Approximation

- Laplace's method

Consider a distribution on a continuous variable of the form

$$p(\mathbf{x}) = \frac{1}{Z} e^{-E(\mathbf{x})}. \quad (28.2.1)$$

The Laplace method makes a Gaussian approximation of $p(\mathbf{x})$ based on a local perturbation expansion around a mode \mathbf{x}^* . First we find the mode numerically, giving

$$\mathbf{x}^* = \underset{\mathbf{x}}{\operatorname{argmin}} E(\mathbf{x}). \quad (28.2.2)$$

Then a Taylor expansion up to second order around this mode gives

$$E(\mathbf{x}) \approx E(\mathbf{x}^*) + (\mathbf{x} - \mathbf{x}^*)^\top \nabla E|_{\mathbf{x}^*} + \frac{1}{2} (\mathbf{x} - \mathbf{x}^*)^\top \mathbf{H} (\mathbf{x} - \mathbf{x}^*) \quad (28.2.3)$$

where $\mathbf{H} \equiv \nabla \nabla E(\mathbf{x})|_{\mathbf{x}^*}$ is the Hessian evaluated at the mode. At the mode, $\nabla E|_{\mathbf{x}^*} = \mathbf{0}$, and an approximation of the distribution is given by the Gaussian

$$q(\mathbf{x}) = \frac{1}{Z_q} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{x}^*)^T \mathbf{H}(\mathbf{x}-\mathbf{x}^*)} = \mathcal{N}(\mathbf{x}|\mathbf{x}^*, \mathbf{H}^{-1}) \quad (28.2.4)$$

which has mean \mathbf{x}^* and covariance \mathbf{H}^{-1} , with $Z_q = \sqrt{\det(2\pi \mathbf{H}^{-1})}$. We can use the above expansion to estimate the integral

$$\int_{\mathbf{x}} e^{-E(\mathbf{x})} \approx \int_{\mathbf{x}} e^{-E(\mathbf{x}^*) - \frac{1}{2}(\mathbf{x}-\mathbf{x}^*)^T \mathbf{H}(\mathbf{x}-\mathbf{x}^*)} = e^{-E(\mathbf{x}^*)} \sqrt{\det(2\pi \mathbf{H}^{-1})}. \quad (28.2.5)$$

The Laplace Gaussian fit to a distribution is not necessarily the ‘best’ Gaussian approximation. As we’ll see below, other criteria, such as those based on minimal KL divergence between $p(\mathbf{x})$ and a Gaussian approximation may be more appropriate, depending on the context. A benefit of Laplace’s method is its relative simplicity compared with other approximate inference techniques.

- So we can approximate the model likelihood with a gaussian distribution

$$\log \Pr(D | M) \approx$$

$$\log \Pr(D | \theta^*, M) + \log \Pr(\theta^* | M) + \frac{1}{2} \log \det(2\pi\mathbf{H}^{-1})$$

where

$$\theta^* = \arg \max_{\theta} \Pr(D | \theta, M) \Pr(\theta | M)$$

and \mathbf{H} is the Hessian of

$$f(\theta) \equiv -\log \Pr(D | \theta, M) \Pr(\theta | M)$$

evaluated at θ^*

- A simpler version of laplace's method

set: $\mathbf{H} = N\mathbf{I}_K$, $K = \dim \theta$

- laplace approximation:

$$\log \Pr(D | M) \approx \log \Pr(D | \theta^*, M) + \log \Pr(\theta^* | M) + \frac{1}{2} \log \det(2\pi\mathbf{H}^{-1})$$

$$\frac{1}{2} \log \det(2\pi\mathbf{H}^{-1}) = \frac{1}{2} \log \det(2\pi \frac{1}{N} \mathbf{I}_K) = \frac{1}{2} \log \left(\frac{2\pi}{N}\right)^K = \frac{K}{2} \log 2\pi - \frac{K}{2} \log N$$

- So $\log \Pr(D | M) \approx \log \Pr(D | \theta^*, M) + \log \Pr(\theta^* | M) + \frac{K}{2} \log 2\pi - \frac{K}{2} \log N$

- More simpler: assume $\Pr(\theta | M) = N(\theta | \mathbf{0}, \mathbf{I})$

$$\log \Pr(\theta^* | M) = \log \frac{1}{(2\pi)^{\frac{K}{2}} |\mathbf{I}_K|^{\frac{1}{2}}} \exp\left[-\frac{1}{2} (\theta^*)^T \mathbf{I}_K \theta^*\right] = -\frac{1}{2} (\theta^*)^T \theta^* - \frac{K}{2} \log 2\pi$$

- Finally we get BIC: $BIC = \log \Pr(D | \theta^*, M) - \frac{K}{2} \log N$



- Bayes' rule enables us to evaluate models based on how well they fit the data via the model likelihood.
- There is no need to explicitly penalise 'complex' models in the Bayesian approach since it automatically incorporates an Occam's razor effect due to the integral over the posterior parameter distribution.
- Computing the model likelihood can be a complex task. In continuous parameter models, Laplace's method provides a simple approximation, the **BIC being a cruder version of Laplace's approximation.**
- Assessing performance on the basis of a limited amount of data can be achieved using simple Bayesian hypothesis testing.



Next

IV: Minimum Description Length

- The basic idea:
 - The goal of statistical inference may be cast as trying to find regularity in the data. 'Regularity' may be identified with 'ability to compress'. MDL combines these two insights by viewing learning as **data compression**: it tells us that, for a given set of hypotheses H and data set D, we should try to find the hypothesis or combination of hypotheses in H that **compresses D most**.

00010001000100010001 ... 0001000100010001000100010001 (1.1)

01110100110100100110 ... 1010111010111011000101100010 (1.2)


00011000001010100000 ... 0010001000010000001000110000 (1.3)

- So MDL has some following properties:

- Occam's Razor
- No overfitting, automatically
- Bayesian interpretation
- No need for 'underlying truth'
- Predictive interpretation

```

for i = 1 to 2500; print '0001'; next; halt
print '011101001101000010101010.....1010111010111011000101100010'; halt
  
```





- Kolmogorov complexity
 - the length of the shortest program that prints the sequence and then halts.
 - the lower, the more regular.
 - leading to an idealized version of MDL
 - Uncomputability
 - Arbitrariness/dependence on syntax

- Hypotheses VS. models
 - we use the phrase *point hypothesis* to refer to a single probability distribution or function. also known as simple hypothesis
 - we use the word *model* to refer to a **family(set)** of probability distribution or functions **with same functional form**. Also known as composite hypothesis.
- An example

hypothesis

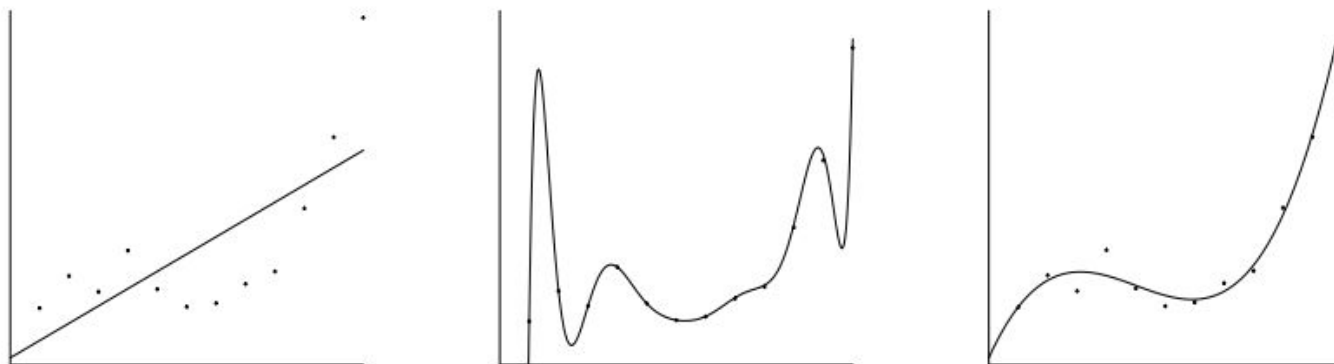


Figure 1.1: A simple, a complex and a trade-off (3rd degree) polynomial.

- Hypothesis selection:
 - if we are interested in selecting both the degree of a polynomial and the corresponding parameters;
- Model selection problem:
 - if we are mainly interested in selecting the degree.

Crude⁴, Two-part Version of MDL Principle (Informally Stated)

Let $\mathcal{H}^{(1)}, \mathcal{H}^{(2)}, \dots$ be a list of candidate models (e.g., $\mathcal{H}^{(k)}$ is the set of k -th degree polynomials), each containing a set of point hypotheses (e.g., individual polynomials). The best point hypothesis $H \in \mathcal{H}^{(1)} \cup \mathcal{H}^{(2)} \cup \dots$ to explain the data D is the one which minimizes the sum $L(H) + L(D|H)$, where

- $L(H)$ is the length, in bits, of the description of the hypothesis; and
- $L(D|H)$ is the length, in bits, of the description of the data when encoded with the help of the hypothesis.

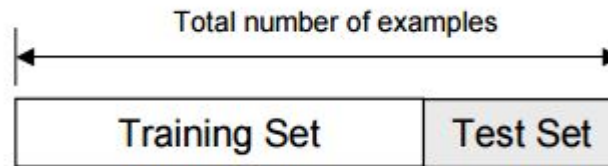
The best *model* to explain D is the smallest model containing the selected H .



Next

Cross-validation

- Hold out method
 - split dataset into two groups
 - training set:used to train the classifier
 - test set:used to estimate the error rate of the trained classifier.

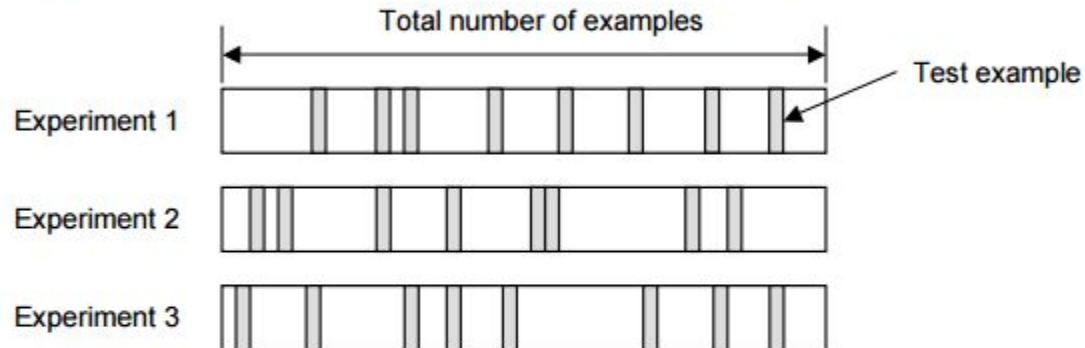


- Cross-validation
 - Random Subsampling
 - K-Fold Cross-Validation
 - Leave-one-out Cross-Validation

- Random Subsampling

- **Random Subsampling performs K data splits of the dataset**

- Each split randomly selects a (fixed) no. examples without replacement
- For each data split we retrain the classifier from scratch with the training examples and estimate E_i with the test examples



- **The true error estimate is obtained as the average of the separate estimates E_i**

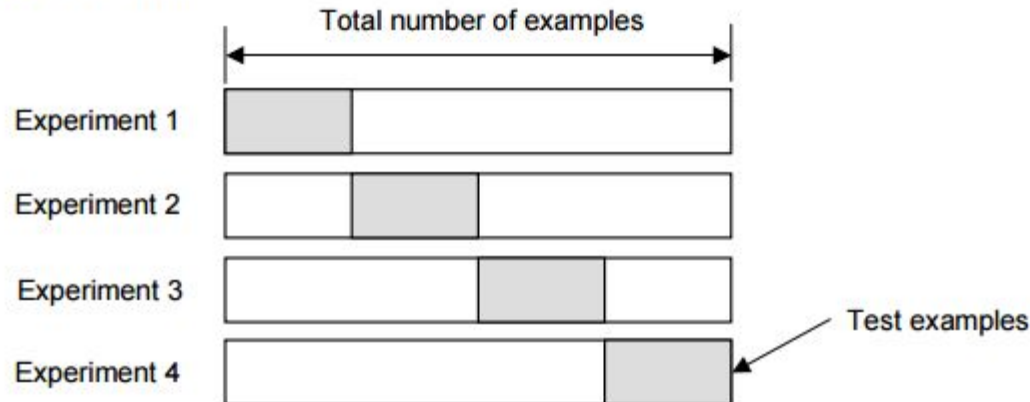
- This estimate is significantly better than the holdout estimate

$$E = \frac{1}{K} \sum_{i=1}^K E_i$$

- K-Fold Cross-Validation

- **Create a K-fold partition of the the dataset**

- For each of K experiments, use K-1 folds for training and the remaining one for testing



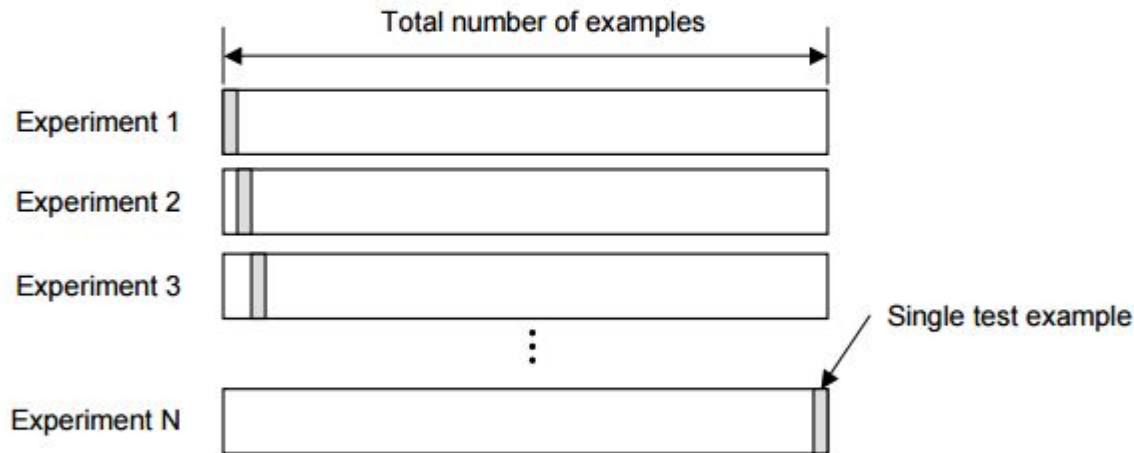
- **K-Fold Cross validation is similar to Random Subsampling**

- The advantage of K-Fold Cross validation is that all the examples in the dataset are eventually used for both training and testing

- **As before, the true error is estimated as the average error rate**

$$E = \frac{1}{K} \sum_{i=1}^K E_i$$

- Leave-one-out Cross-Validation
 - **Leave-one-out is the degenerate case of K-Fold Cross Validation, where K is chosen as the total number of examples**
 - For a dataset with N examples, perform N experiments
 - For each experiment use N-1 examples for training and the remaining example for testing



- **As usual, the true error is estimated as the average error rate on test examples**

$$E = \frac{1}{N} \sum_{i=1}^N E_i$$



Next

V:Others



- Deviance information criterion
- False discovery rate
- Focused information criterion
- Mallows's C_p
- Minimum message length (Algorithmic information theory)
- Structural Risk Minimization
- Stepwise regression



- [wiki explanation](#)
- [Thoughts on model selection](#)(Richard A. Davis Colorado State University)
- [Bayesian Reasoning and Machine learning\(chapter 12\)](#)
- [lecture about model selection](#)