《数据挖掘技术》

⭐ CH01 统计学习方法概论

➡️ Create by *Wang JingHui*

➡️ Version *4.0*

主要内容

1. 统计学习
2. 统计学习分类
3. 统计学习三要素 - 模型/策略/算法
4. 模型评估与模型选择
5. 正则化与交叉验证
6. 泛化能力
7. 生成模型与判别模型
8. 监督学习应用

统计学习

统计学习的对象

- 数据： 计算机及互联网上的各种数字、文字、图像、视 频、音频数据以及它们的组合。
- 数据的基本假设是同类数据具有**一定的统计规律性**。

统计学习的目的

- 用于对数据（特别是未知数据）进行预测和分析。

统计学习的分类

基本分类

- 监督学习
- 非监督学习
- 半监督学习
- 强化学习

按模型分类

- 概论模型与非概论模型
- 线性模型与非线性模型
- 参数化模型与非参数化模型

按算法分类

- 在线学习
- 批量学习

按技巧分类

- 贝叶斯学习
- 核方法

统计学习方法三要素

统计学习方法 = 模型 + 策略 + 算法

> 1. 得到一个有限的训练数据集合
> 2. 确定包含所有可能的模型的**假设空间**, 即学习模型的集合.
> 3. 确定模型选择的准则, 即学习的**策略**
> 4. 实现求解最优模型的算法, 即学习的**算法**
> 5. 通过学习方法选择最优的模型
> 6. 利用学习的最优模型对新数据进行预测或分析.

模型是什么?

在监督学习过程中, **模型**就是所要学习的**条件概率分布**或者**决策函数**.

- 模型表达了输入到输出的一种映射集合，这个集合就是假设空间，假设空间表明模型学习的范围。
- 假设空间包含了所有可能的条件概率分布或决策函数；

- 假设空间用$\mathcal{F}$表示，假设空间可以定义为决策函数的集合：

$$\mathcal{F} = \{f|Y = f(X)\}$$

其中，X和Y是定义在输入空间$\mathcal{X}$和输出空间$\mathcal{Y}$上的变量，$\mathcal{F}$通常是由一个参数向量决定的函数族：

$$\mathcal{F} = \{f|Y = f_\theta(x), \theta \in R^n\}$$

参数向量θ取值于n为欧式空间R^n，称为**参数空间**。

- 假设空间也可以定义为条件概率的集合

$$\mathcal{F} = \{P|P(Y|X)\}$$

其中，X和Y是定义在输入空间$\mathcal{X}$和输出空间$\mathcal{Y}$上的变量，$\mathcal{F}$条件概率分布族：

$$\mathcal{F} = \{P|P_\theta(Y|X)\}, \theta \in R^n\}$$

参数向量θ取值于n为欧式空间R^n，称为**参数空间**。

总结

	假设空间$\mathcal{F}$	输入空间$\mathcal{X}$	输出空间$\mathcal{Y}$	参数空间
决策函数	$\mathcal{F}=\{f_\theta \vert Y=f_\theta(x),\theta\in\mathbf{R}^n\}$	变量	变量	$\mathbf{R}^n$
条件概率分布	$\mathcal{F}=\{P \vert P_\theta(Y\vert X),\theta\in\mathbf{R}^n\}$	随机变量	随机变量	$\mathbf{R}^n$

由决策函数表示的模型为非概率模型；
由条件概率表示的模型为概率模型。

策略

损失函数与风险函数

> **损失函数**度量模型**一次预测**的好坏，**风险函数**度量**平均意义**下模型预测的好坏。

- 损失函数(loss function)或代价函数(cost function)
  损失函数定义为给定输入X的**预测值**$f(X)$**和真实值**Y**之间的非负实值**函数, 记作$L(Y, f(X))$

常用损失函数 -数值越小，模型就越好。

1. 0-1损失

$$L=\begin{cases}1, Y \neq f(X) \\ 0, Y=f(X)\end{cases}$$

2. 平方损失

$$L=(Y-f(X))^2$$

3. 绝对损失

$$L=|Y-f(X)|$$

4. 对数损失

$$L(Y, P(Y|X))=-\log P(X|Y)$$

- $P(Y|X) \leqslant 1$，对应的对数$L=-\log P(Y|X)$是负值。

风险函数(risk function)或期望损失(expected loss)

这个和模型的泛化误差的形式是一样的

$$R_{exp}(f) = E_p[L(Y, f(X))] = \int_{\mathcal{X}\times\mathcal{Y}} L(y, f(x))P(x, y)\,\mathrm{d}x\mathrm{d}y$$

模型$f(X)$关于联合分布$P(X, Y)$的**平均意义下的**损失(**期望**损失), 但是因为$P(X, Y)$是未知的, 所以前面的用词是**期望**, 以及**平均意义下的**.

这个表示其实就是损失的均值, 反映了对整个数据的预测效果的好坏, $P(x, y)$转换成$\frac{\nu(X=x, Y=y)}{N}$更容易直观理解, 真实的数据N是无穷的.

经验风险最小化与结构风险最小化

1. **经验风险**(empirical risk)或**经验损失**(empirical loss)

$$R_{emp}(f) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i))$$

模型f关于**训练样本集**的平均损失

根据大数定律, 当样本容量N趋于无穷大时, 经验风险趋于期望风险

2. **结构风险**(structural risk)

$$R_{srm}(f) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f)$$

$J(f)$为模型复杂度, $\lambda \geqslant 0$是系数, 用以权衡经验风险和模型复杂度.

ERM与SRM

经验风险最小化(ERM)与结构风险最小化(SRM)

1. **极大似然估计**是经验风险最小化的一个例子.
   当模型是条件概率分布, 损失函数是对数损失函数时, 经验风险最小化等价于极大似然估计.
2. **贝叶斯估计**中的**最大后验概率估计**是结构风险最小化的一个例子.
   当模型是条件概率分布, 损失函数是对数损失函数, **模型复杂度由模型的先验概率表示**时, 结构风险最小化等价于最大后验概率估计.

算法

统计学习归结为最优化问题。

- 如果最优化问题有显式的解析解，直接使用；
- 解析解不存在，需要数值计算的方法。

总结

统计方法之间的不同，主要来自其模型、策略和算法的不同。

模型评估与模型选择

训练误差与测试误差

统计学习的目的是得到的模型不仅对已知数据而且对未知数据都能有很好的预测能力。

训练数据集的平均损失

$$R_{emp}(\hat{f}) = \frac{1}{n}\sum_{i=1}^{N} L(y_i, \hat{f}(x_i))$$

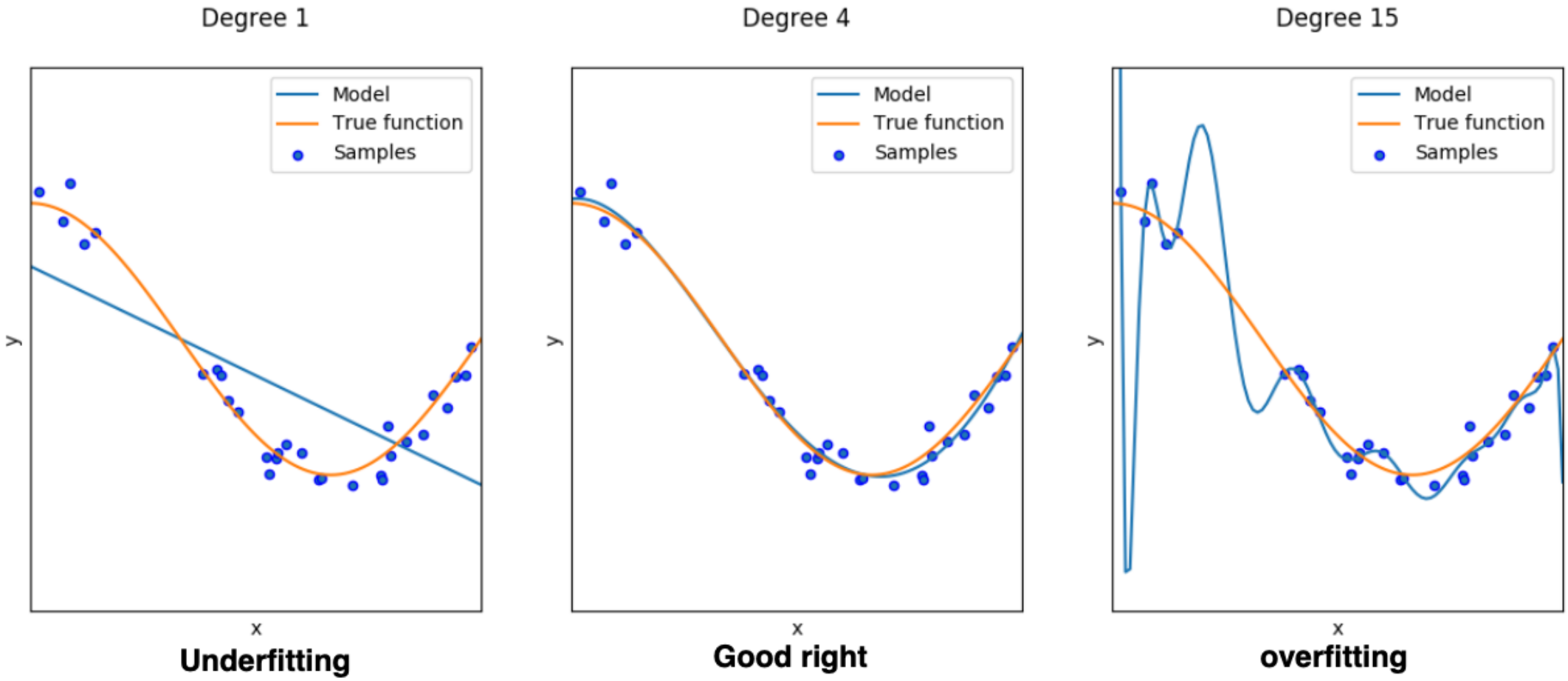
测试数据集的平均损失

$$R_{test}(\hat{f}) = \frac{1}{n}\sum_{i=1}^{N} L(y_i, \hat{f}(x_i))$$

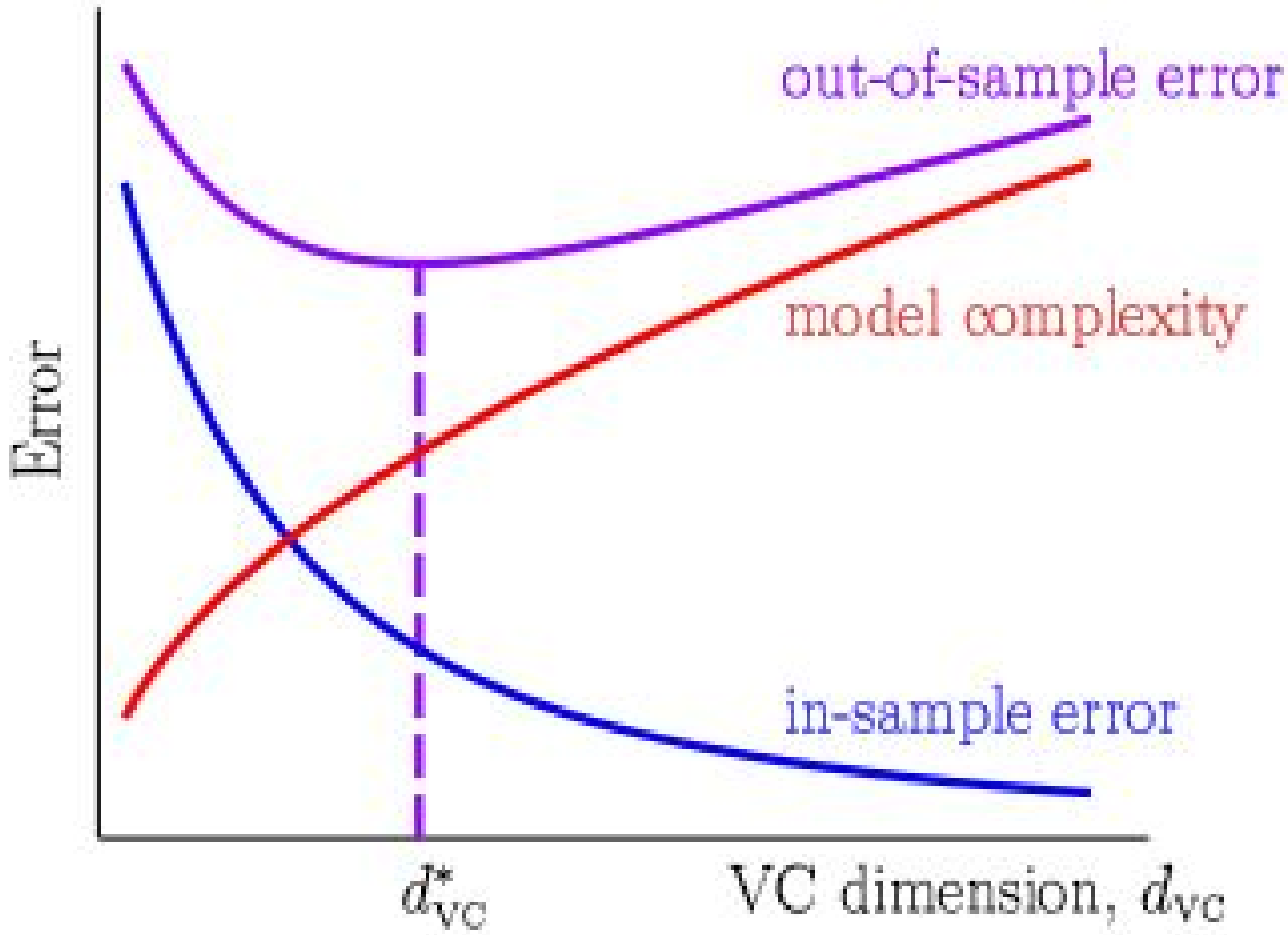
- 当损失函数是0-1损失时：$R_{test} = \frac{1}{n}\sum_{i=1}^{N} I(y_i \neq \hat{f}(x_i))$

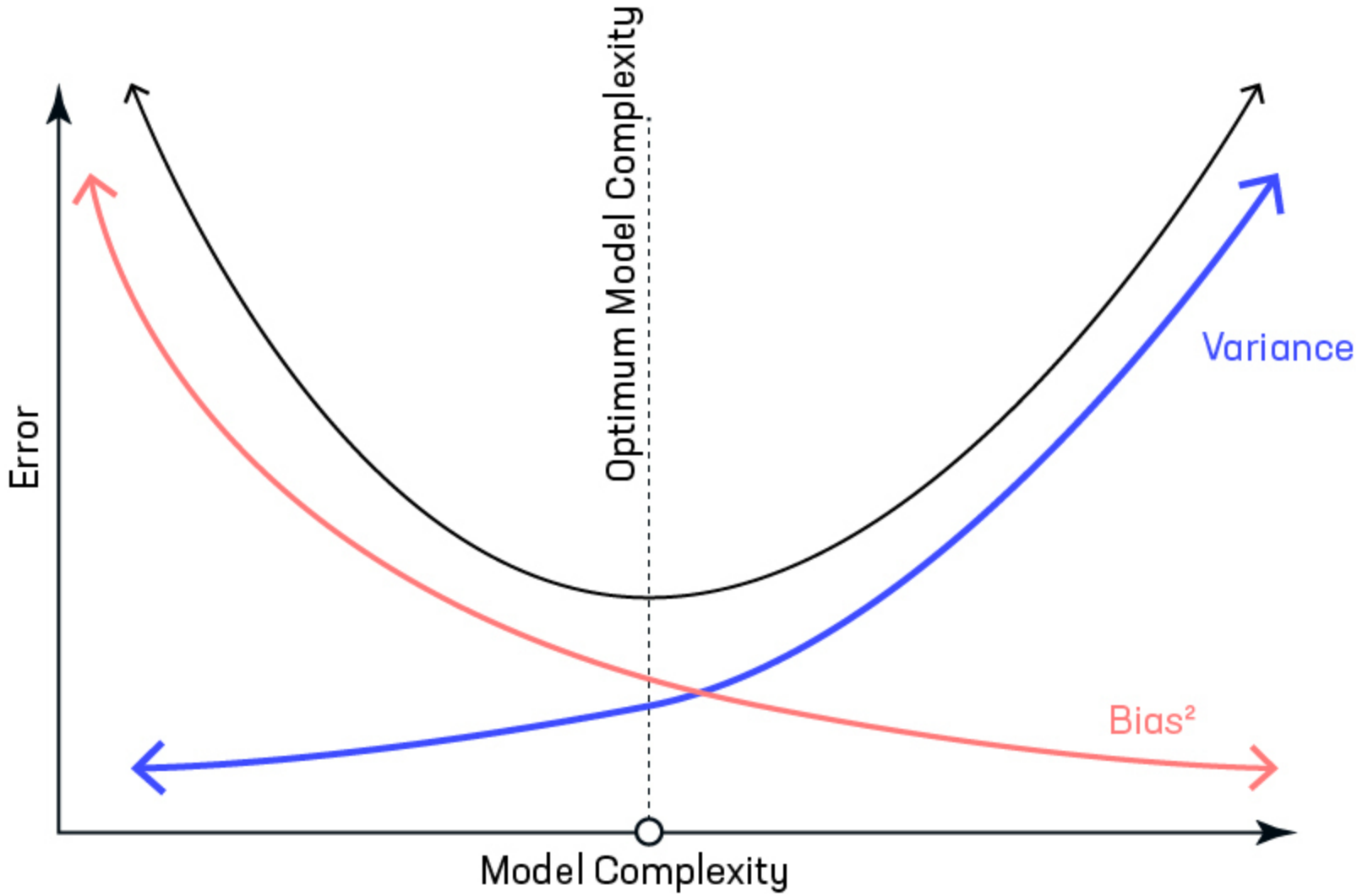
过拟合与模型选择

使用M次多项式函数中选择一个对已知数据以及未知数据都有很好预测能力的函数。

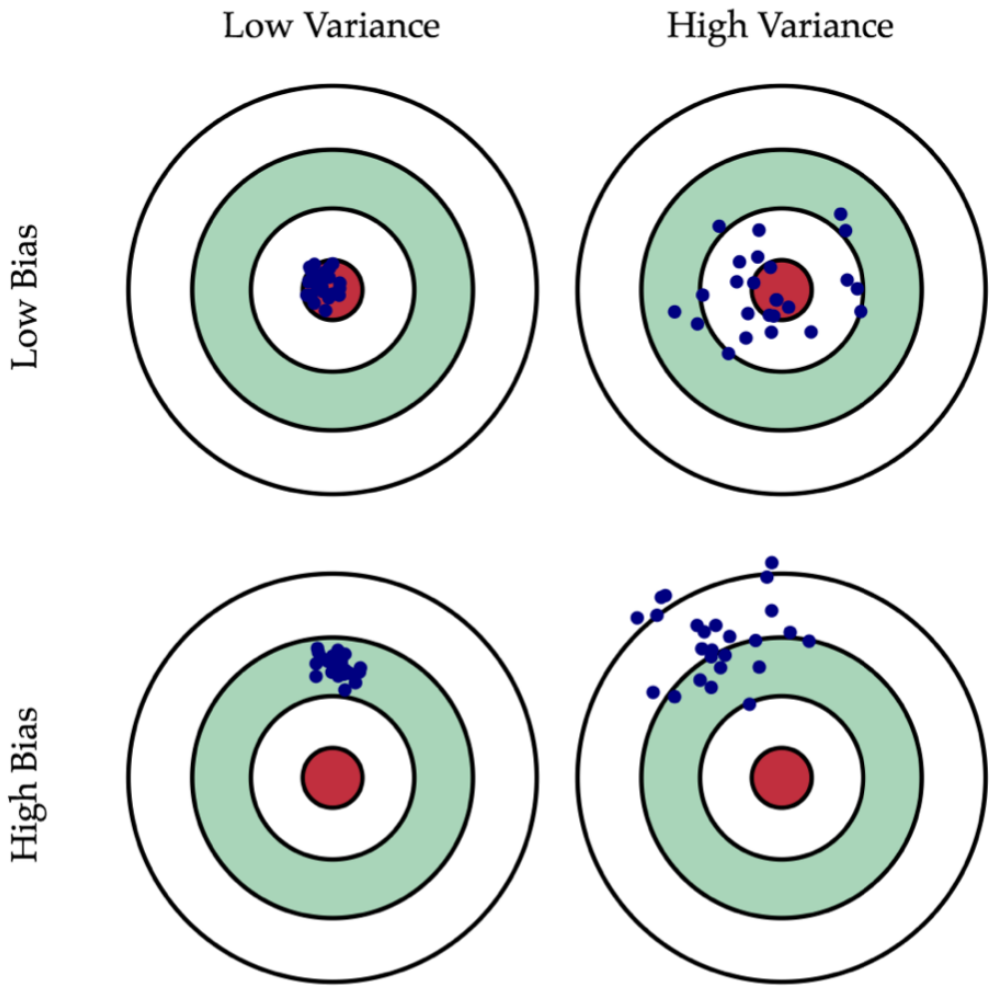


Overfitting and Model Selection




Optimum Model Complexity
Error
Variance
Bias²
Model Complexity

Variance VS Bias



There are 4 cases representing combinations of both high and low bias and variance.

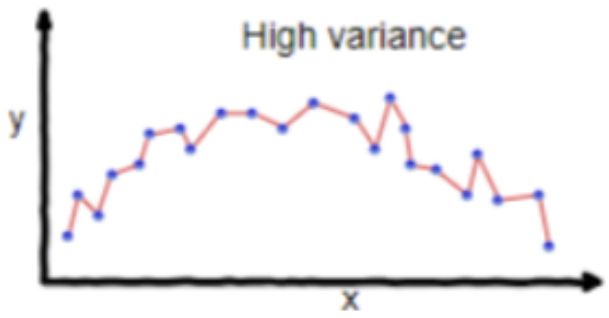
- Low Bias & Low Variance → Good case
- Low Bias & High Variance
- High Bias & Low Variance
- High Bias & High Variance → Bad case

Ideally, one wants to choose a model that both accurately captures the regularities in its training data, but also **generalize**s well to unseen data (i.e. new data). Unfortunately, it is typically impossible to do both simultaneously[1].

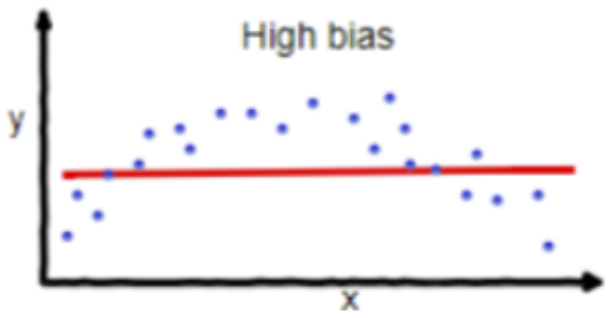
Model Selection

- There is usually a trade-off between Bias and Variance. So normally reducing one tends to increase the other
- Select a model that balances two kinds of error to minimize the total error

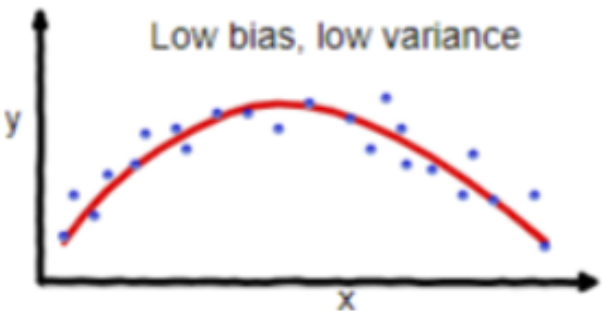
Variance VS Bias



overfitting



underfitting



Good balance

正则化与交叉验证

正则化

模型选择的典型方法是**正则化**。

- 正则化是结构风险最小化策略的实现，在**经验风险**上加一个**正则化项(regularier)或惩罚项(penalty term)**。
- 正则化项一般是模型复杂度的单调递增函数，模型越复杂，正则化值就越大；
- 正则化的一般形式：

$$\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f)$$

例： 回归问题中，选择$H(x)$为n次的多项式，$H(x)=w_0+w_1x+w_2x^2+...w_nx^n$
其中$w(w_0,w_1,w_2,...,w_n)$为参数；

则问题可以表示为：

$$L(w)=\frac{1}{N}\sum_{i1}^{N}(f(x_i;w)-y_i)^2+\frac{\lambda}{2}||w||^2$$

$$L(w)=\frac{1}{N}\sum_{i1}^{N}(f(x_i;w)-y_i)^2+\lambda||w||_1$$

$\lambda(f)$ 模型复杂度

L0范数-表示向量中非零元素的个数：

$$||w||_0 = \#(i), w_i \neq 0$$

- 即希望w中的大部分元素都是0，（w是稀疏的），用于稀疏编码，最优化问题是一个NP问题，L1范数是L0范数的最优凸近似，因此通常使用L1范数来代替。

L1范数--(Lasso Regression)-表示向量中每个元素绝对值的和：

$$||w||_1 = \sum_{i=1}^{n} |w_i|$$

- L1范数的解通常是稀疏性的，倾向于选择数目较少的一些非常大的值或者数目较多的insignificant的小值。

L2范数--(Ridge Regression)-欧氏距离：

$$||w||_2 = \sqrt{\sum_{i=1}^{n} w_i^2}$$

L2范数越小，可以使得w的每个元素都很小，接近于0，但L1范数不同的是他不会让它等于0而是接近于0.

例： 列向量$[3, 0, 0, 0, -4]^T$，求$L0, L1, L2$。

奥卡姆剃刀

- 在所有可能选择的模型中，能够很好地解释已知数据并且十分简单才是最好的模型，也就是应该选择的模型。

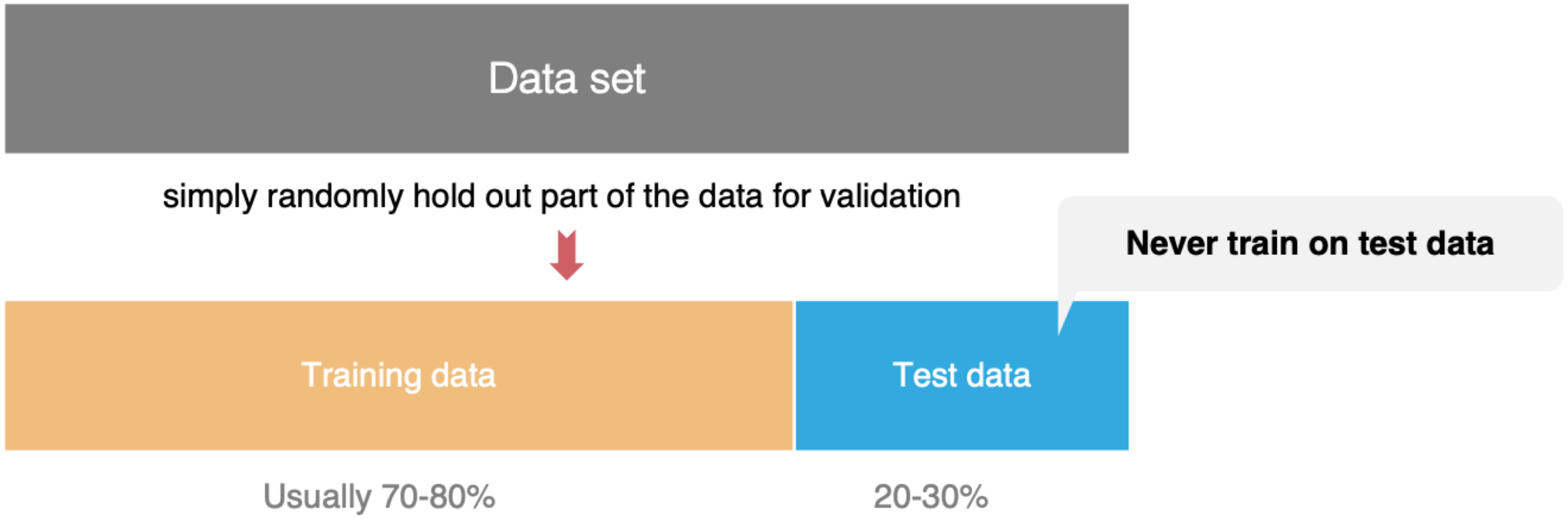
交叉验证

交叉验证是常用的模型选择方法。

数据集分为：训练集、验证集、测试集

Model validation strategy : Hold-out validation

- 简单交叉验证

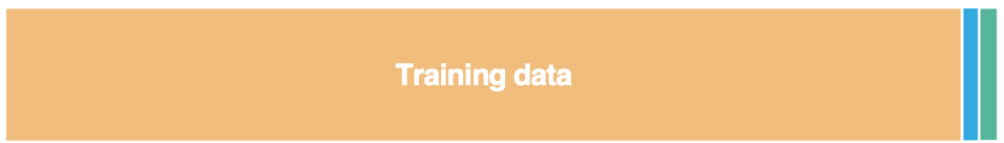


simply **randomly** hold out part of the data for cross validation & test

For relatively small data set:

Usually 60% 20% 20%

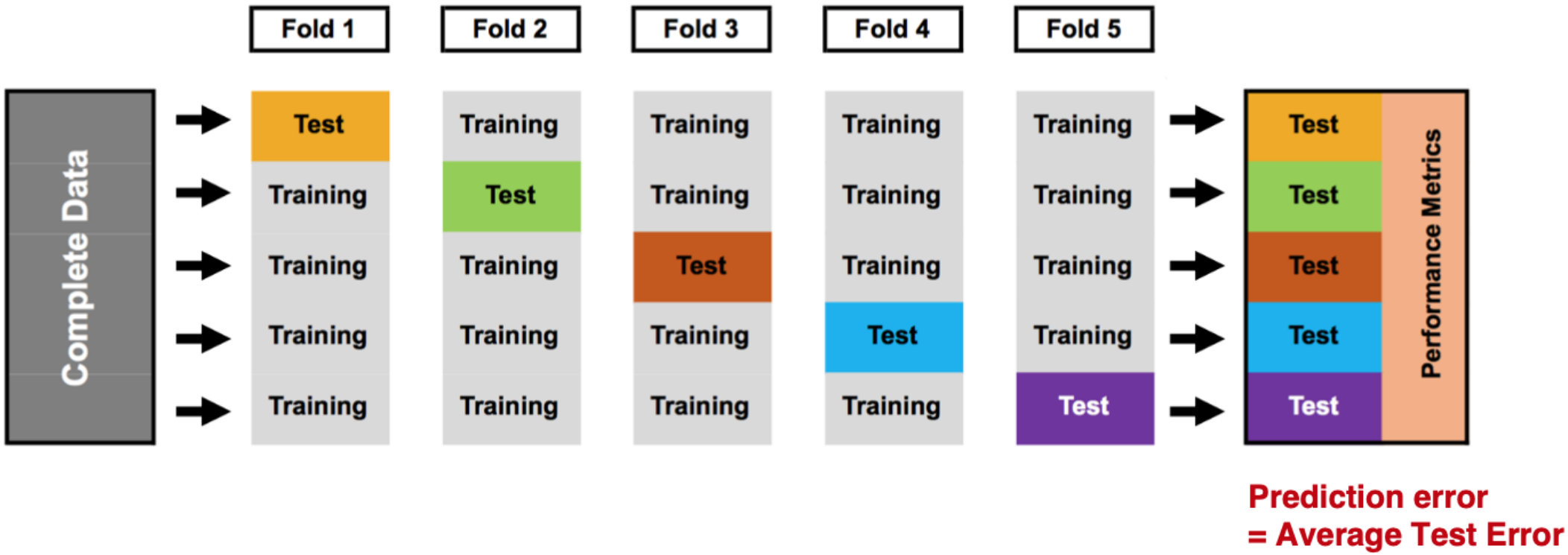
For big data set:



98% 1% 1%

Model validation strategy : k-fold cross validation

- S折交叉验证(K折, K-Fold cross validation)[^1]
  - 先将数据集D划分为k个大小相似的互斥子集，即$D = D_1 \cup D_2 \cup ... \cup D_k$，$D_i \cap D_j$，每个子集$D_i$都尽可能保持数据分布的一致性，即从$D$中通过分层采样得到，用$k-1$个子集的并集作为训练集，余下的那个子集作为测试集；从而可以进行k次训练和测试，最终返回的是这k个测试结果的均值。



Prediction error
= Average Test Error

Leave-one-out cross validation:

- specific case of k-fold cross validation, where k = n
- 留一法(Leave One Out,LOO)：S折交叉验证的特殊情形$S = N$

- use one observation as the validation set and and the remaining observations as the training set
- This is repeated such that each observation in the sample is used once as the validation data.
- LOOCV is computationally intensive !

泛化能力与模型选择

- 现实中采用最多的方法是通过测试误差来评价学习方法的泛化能力，统计学习理论试图从理论上对学习方法的泛化能力进行分析。
- 学习方法的泛化能力往往是通过研究泛化误差的**概率上界**进行的, 简称为泛化误差上界(generalization error bound)。
- 注意泛化误差的定义，**事实上，泛化误差就是所学习到的模型的期望风险。**
- 一种方法学习的模型比另一种方法学习的模型具有更小的泛化误差，则这种方法就更有效。
- 泛化误差的性质：样本容量增加，泛化误差趋于0；假设空间容量越大，模型就越难学，泛化误差上界就越大。

例

定理1.1（泛化误差上界） 对二分类问题，当假设空间是有限个函数的集合

$$\mathcal{F} = \{f_1, f_2, ..., f_d\}$$

对任意一个函数$f \in \mathcal{F}$，至少以概率$1 - \delta$，以下不等式成立：

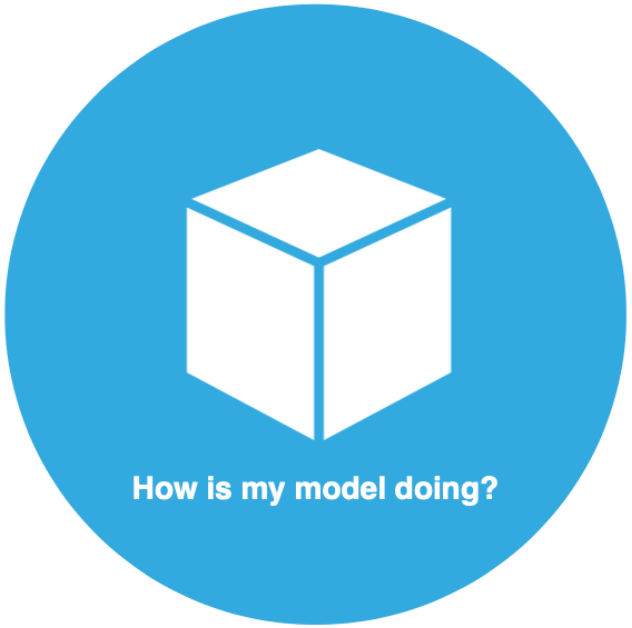
$$R(f) \leq \hat{R}(f) + \epsilon(d, N, \delta)$$

其中

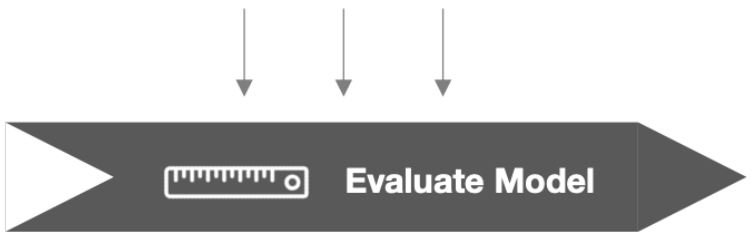
$$\epsilon(d, N, \delta) = \sqrt{\frac{1}{2N}(\log d + \log \frac{1}{\delta})}$$

左边是泛化误差，右边是泛化误差上界。

模型选择



Test Dataset



We need criteria or metrics to evaluate the performance of model

- Accuracy
- Precision
- Recall
- F score
- ROC
- AUC
- Log Loss
- MAE (Mean absolute error)
- MSE (Mean Squared Error)
- RMSE (Root means squared error)
- MAPE (Mean absolute % error)
- R^2 (Coefficient of determination)

Model Evaluation Metrics

Classification Model

- Accuracy
- Precision
- Recall
- F score
- ROC
- AUC
- Log Loss

Regression Model

- MAE (Mean absolute error)
- MSE (Mean Squared Error)
- RMSE (Root means squared error)
- MAPE (Mean absolute % error)
- R^2 (Coefficient of determination)

分类问题模型评价

分类准确率：对于给定的测试数据集，分离器正确分类的样本数与总样本数之比。

通常以关注的类为正类，其他类为负类。

- TP——将正类预测为正类；
- FN——将正类预测为负类；
- FP——将负类预测为正类；
- TN——将负类预测为负类。

混淆矩阵

	预测值 正	预测值 负
真实值 正	TP	FN
真实值 负	FP	TN

精确率(查准率)

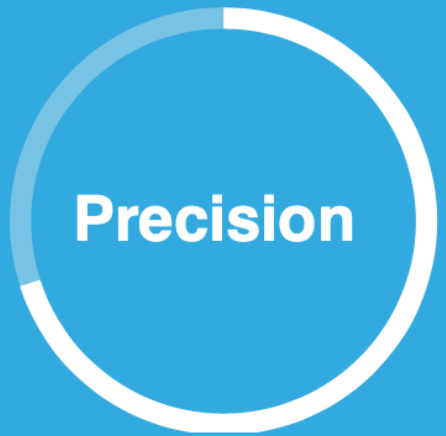
- 精确率是针对我们预测结果而言的，它表示的是预测为正的样本中有多少是真正的正样本。那么预测为正就有两种可能了，一种就是把正类预测为正类(TP)，另一种就是把负类预测为正类(FP)，也就是

$$P = \frac{TP}{TP + FP}$$

召回率(查全率)

- 召回率是针对我们原来的样本而言的，它表示的是样本中的正例有多少被预测正确了。那也有两种可能，一种是把原来的正类预测成正类(TP)，另一种就是把原来的正类预测为负类

$$R = \frac{TP}{TP + FN}$$



Precision attempts to answer the following question:

What proportion of positive identifications was actually correct? [1]

$$precision = \frac{\#\ True\ Positive}{\underbrace{\#\ True\ Positive + \#\ False\ Positive}}$$

No. of <u>predicted</u> positive

Note that the denominator is the count of all data points predicted as positive, including positive observations which were in fact negative.

Recall attempts to answer the following question:

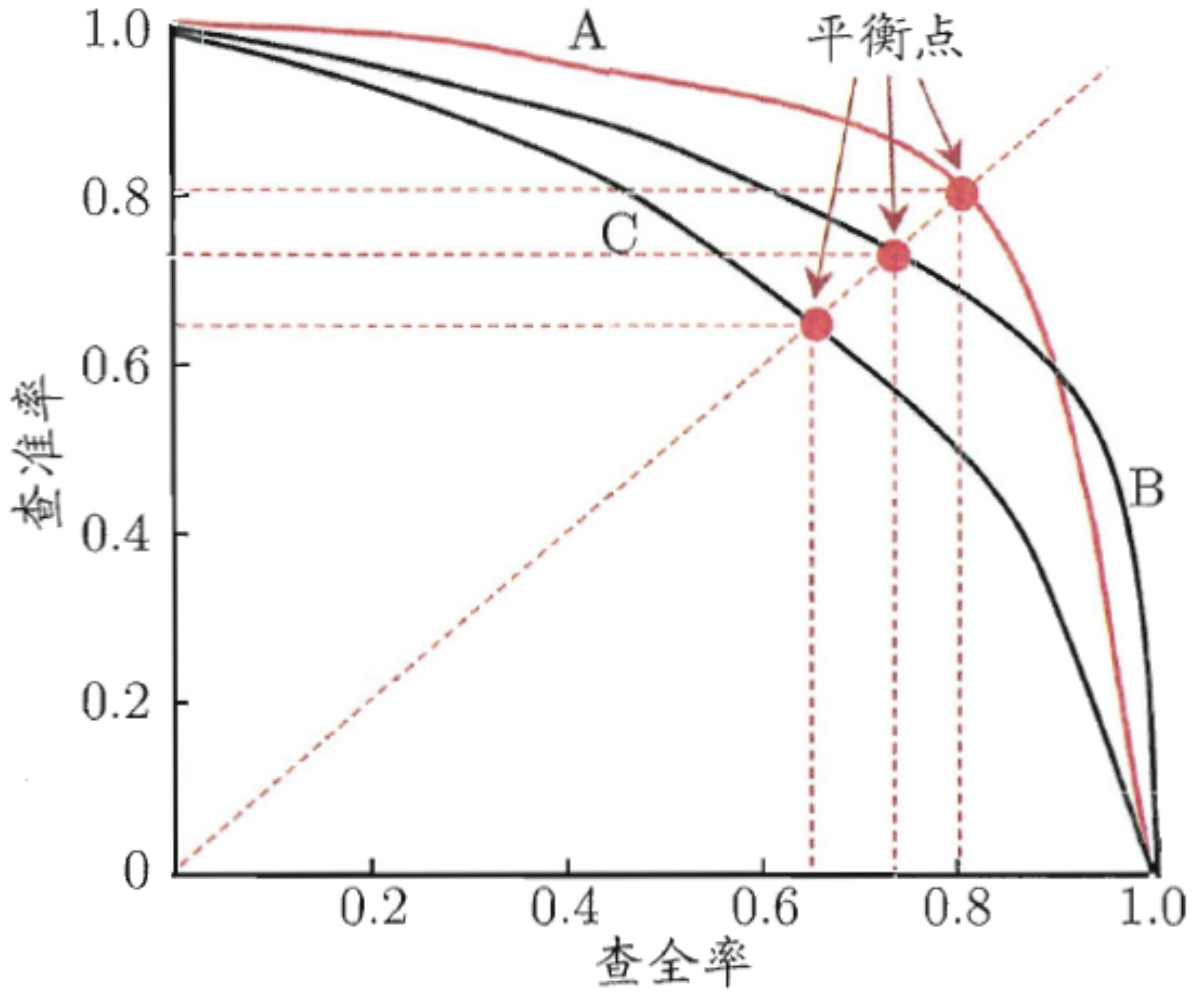
What proportion of actual positives was identified correctly? [2]

$$recall = \frac{\#\ True\ Positive}{\underbrace{\#\ True\ Positive + \#\ False\ Negative}}$$

No. of <u>actual</u> positive

Note that the denominator is the count of all actual/real positive data points. They are all the real positive classes although some of them might be misclassified as negative.

PR曲线



F_1 值是精确率和召回率的调和均值

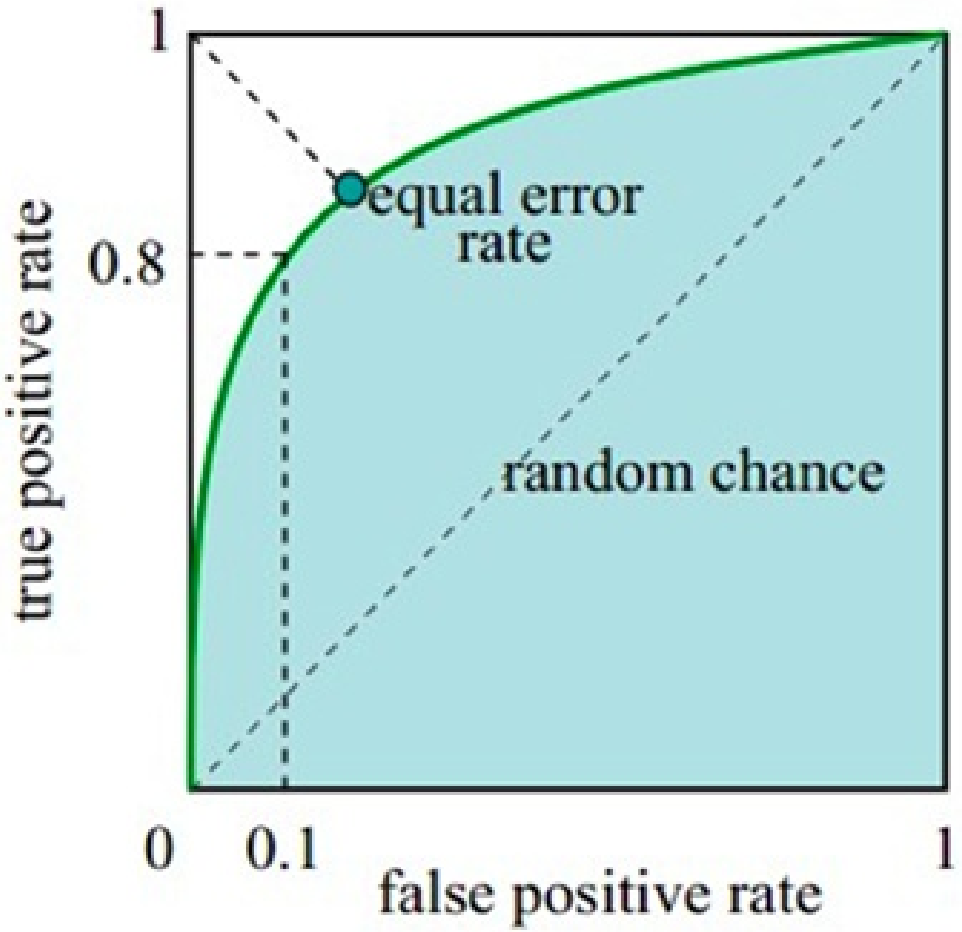
$$\frac{2}{F_1} = \frac{1}{P} + \frac{1}{R}$$

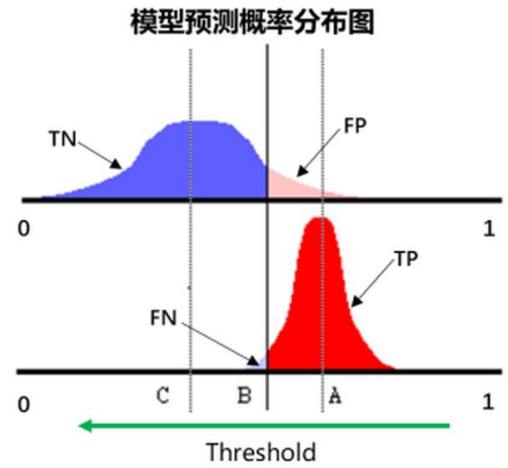
$$F_1 = \frac{2TP}{2TP + FP + FN}$$

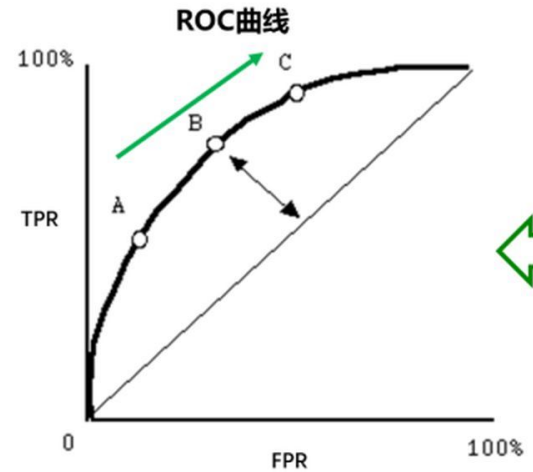
精确率和召回率都高时，F_1 值也会高。

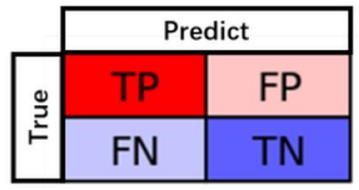
ROC曲线和AUC

- ROC的全称是Receiver Operating Characteristic Curve，中文名字叫“受试者工作特征曲线”，顾名思义，其主要的分析方法就是画这条特征曲线。
- 横坐标：伪正类率(False positive rate， FPR)，预测为正但实际为负的样本占所有负例样本的比例；
- 纵坐标：真正类率(True positive rate， TPR)，预测为正且实际为正的样本占所有正例样本的比例。
- TPR=1，FPR=0，即图中(0,1)点，故ROC曲线越靠拢(0,1)点，越偏离45度对角线越好，Sensitivity、Specificity越大效果越好。







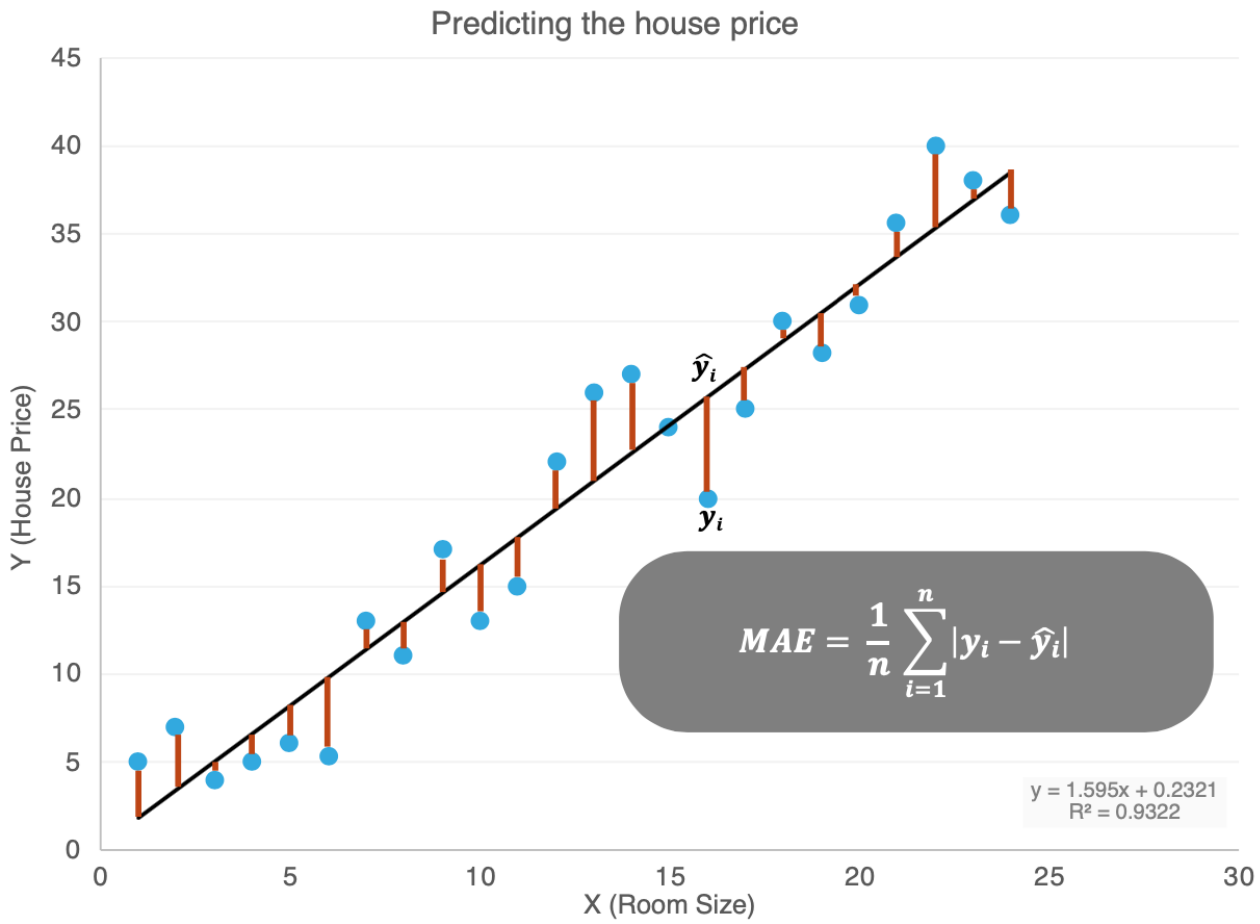


$$\mathrm{TPR} = \frac{\mathrm{TP}}{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

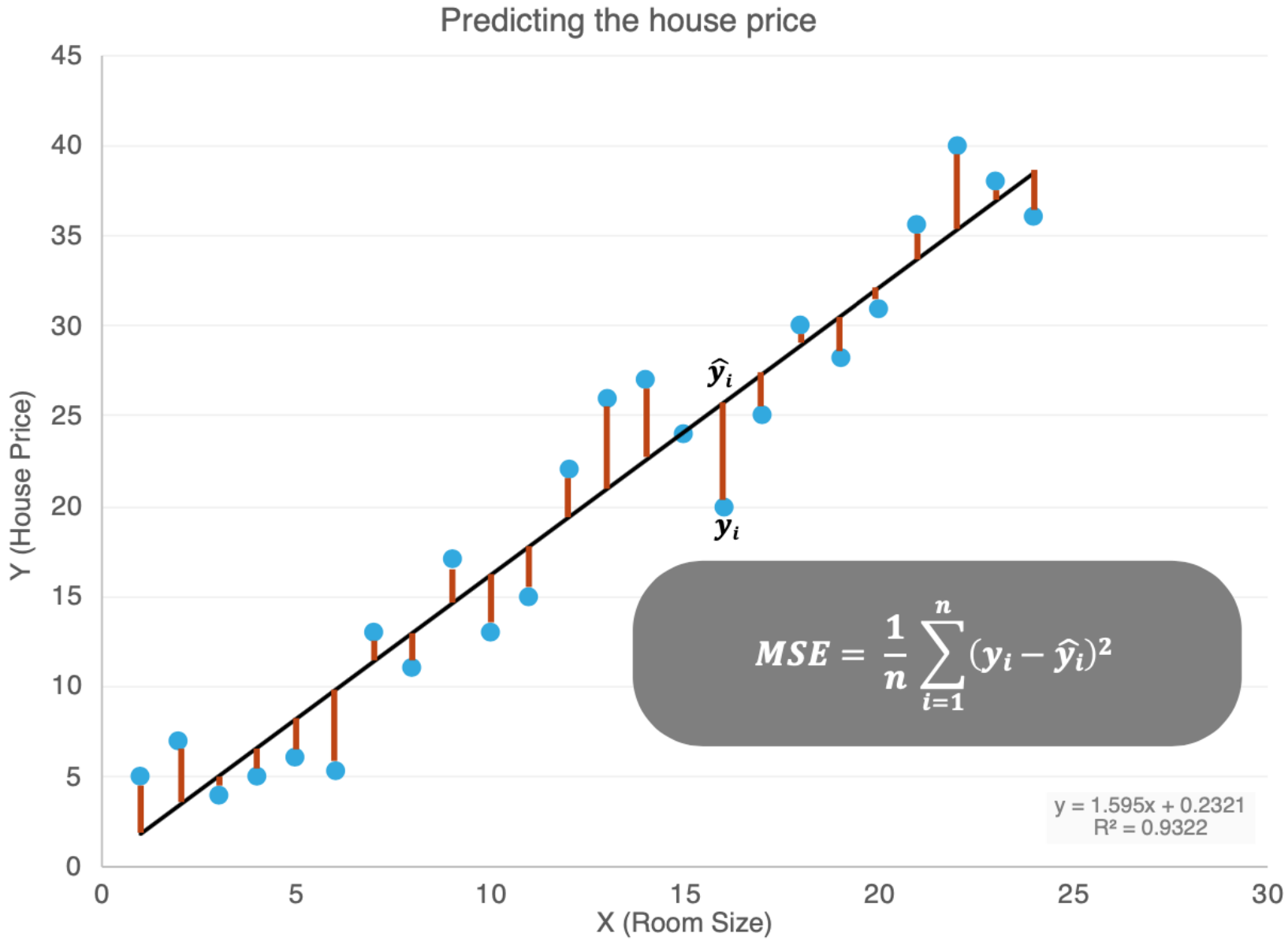
$$\mathrm{FPR} = \frac{\mathrm{FP}}{N} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$$

回归问题模型评价

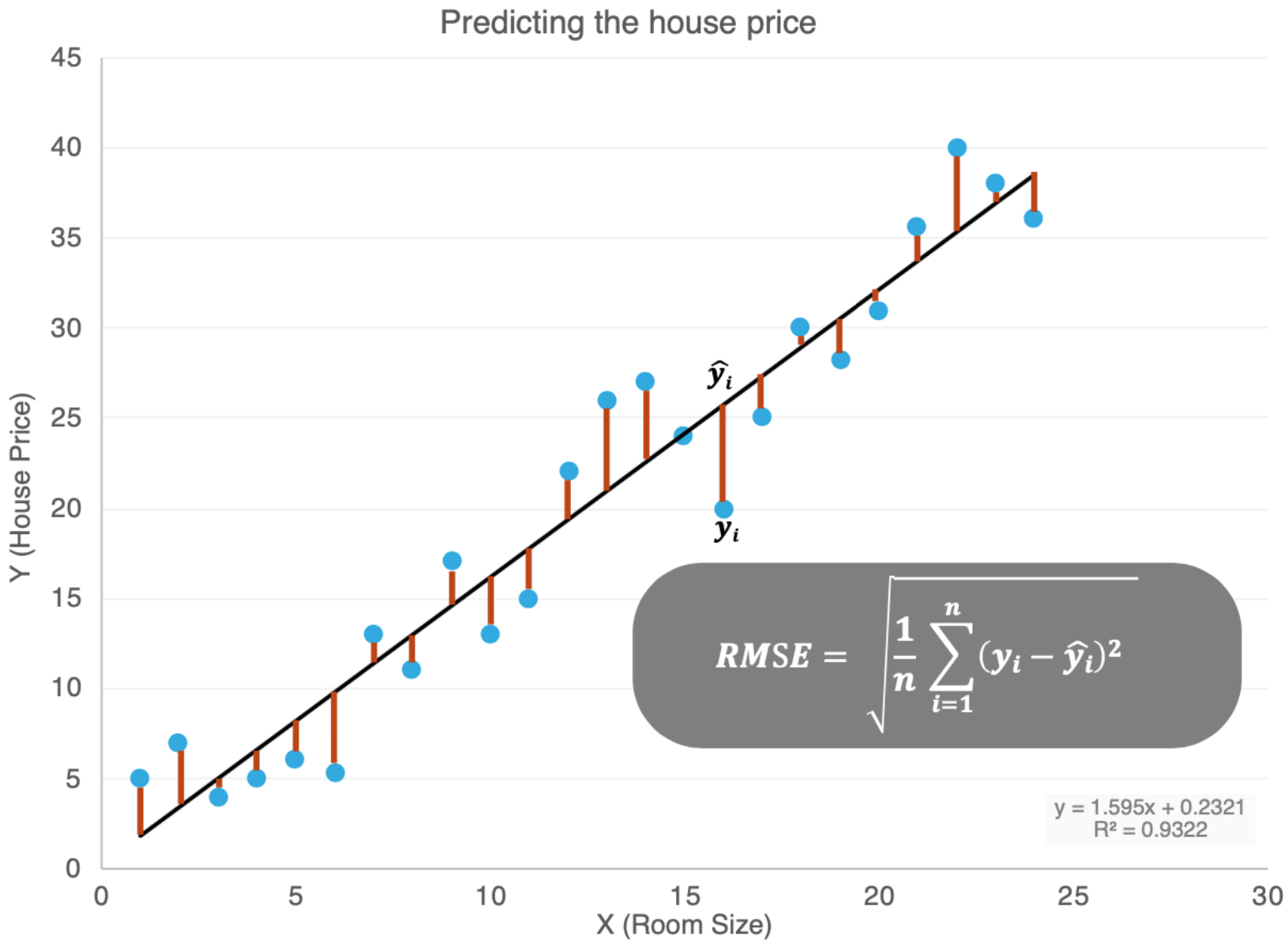
MAE is the mean of the absolute value of the errors



MSE is the mean of the squared value of the errors



RMSE is the square root of the mean of the squared errors



生成模型与判别模型

生成方法 - generative approach

- 可以还原出**联合概率分布**$P(X,Y)$
- 收敛速度快, 当样本容量增加时, 学到的模型可以更快收敛到真实模型
- 当存在隐变量时仍可以用

判别方法 - discriminative approach

- 直接学习**条件概率**$P(Y|X)$或者**决策函数**$f(X)$
- 直接面对预测, 往往学习准确率更高
- 可以对数据进行各种程度的抽象, 定义特征并使用特征, 可以简化学习问题

监督学习应用

分类问题(Classification)

- 监督学习中，当输出变量Y取有限个离散值时，预测问题就成为分类问题；输入变量X可以是离散的，也可以是连续的；
- 从数据中学习得到一个分类模型或分类决策函数，称为分离器；
- 分类问题包含学习和分类两个过程；

标注问题(Tagging)

- 标志问题是分类问题的一个推广；
- 标记问题的输入是一个观测序列，输出是一个标记序列或状态序列；
- 目标是学习一个模型，能够对观测序列给出标记序列作为预测；
- 标注问题分为学习和标注两个过程；
- 常用的统计学习方法有：隐马尔可夫模型、条件随机场。

回归问题(Regression)

- 按照输入变量的个数，分为一元回归和多元回归；
- 按照输入变量和输出变量的关系（模型类型），分为线性回归和非线性回归；
- 最常用的损失函数是平方损失函数，著名的最小二乘法。

参考

[1] ESL,

[2] PRML

Enjoy your machine learning!

https://github.com/wjssx/

E-mail: csr_dsp@sina.com