逻辑回归

机器学习研究室

计算机科学与技术学院

吉林大学

大纲

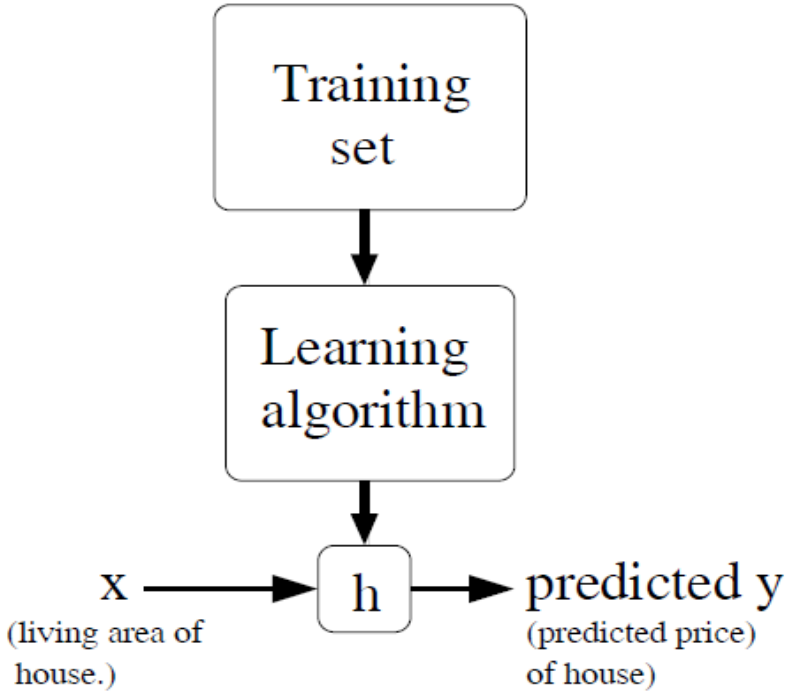
- 分类问题
- Logistics回归模型
- 随机梯度上升算法
- 逻辑回归与感知器的关系
- 牛顿法求解极大似然目标函数
- 线性回归的概率解释
- 模型评估方法和性能评价指标
- 应用实例：垃圾邮件过滤

分类问题

分类问题

- 特征 $x^{(i)}$
- 目标 $y^{(i)}$
- 训练样本 $(x^{(i)}, y^{(i)})$
- 训练集

  $\{(x^{(i)}, y^{(i)}); i = 1, \ldots, m\}$
- 假设 $h(x)$



y 取离散值，是分类问题

分类问题

- 若 y={1，2，3，…,M}，则称此问题为**M分类问题**
- 通常处理的都是二分类问题
- 多分类问题常被转化为多个二分类问题
- 生活中常见的分类应用：
  - 人脸识别
  - 指纹识别
  - 手写体数字识别
  - 垃圾邮件检测

分类问题

- 大型多类图像数据集—ImageNet

IMAGENET

14,197,122 images, 21841 synsets indexed

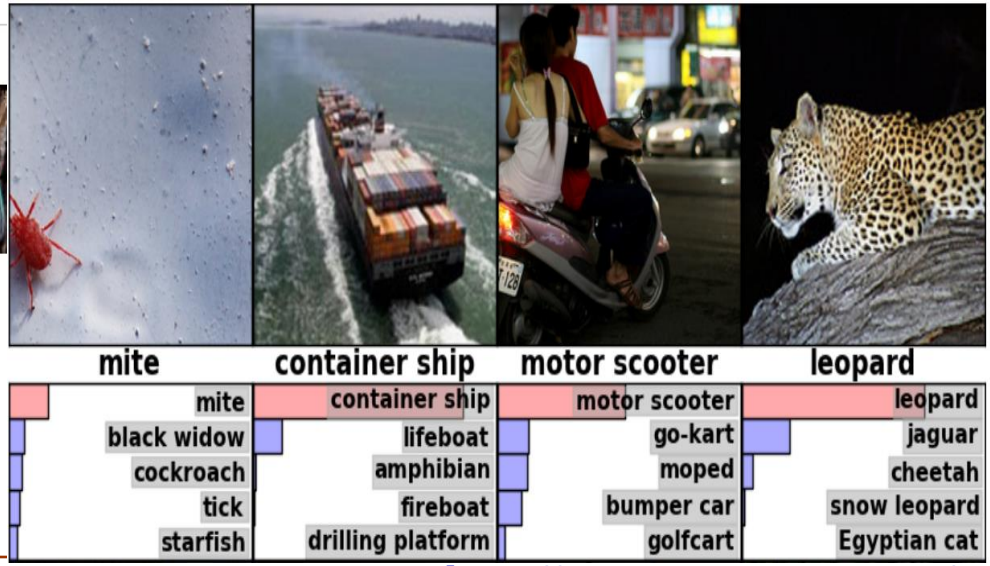
Explore Download Challenges Publications Updates About

Not logged in. Login | Signup

ImageNet is an image database organized according to the WordNet hierarchy (currently only the nouns), in which each node of the hierarchy is depicted by hundreds and thousands of images. Currently we have an average of over five hundred images per node. We hope ImageNet will become a useful resource for researchers, educators, students and all of you who share our passion for pictures.

Click here to learn more about ImageNet, Click here to join the ImageNet mailing list.

- ✓ 14M images
- ✓ 21k synsets indexed



WorldNet hierarchy

https://wordnet.princeton.edu/

http://www.image-net.org/

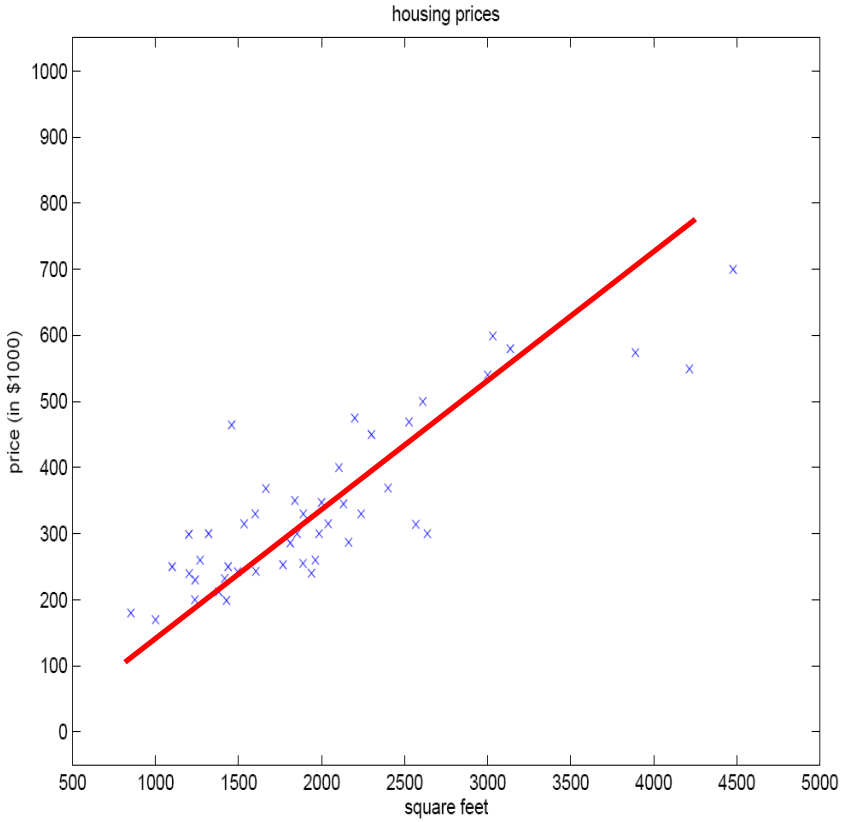
LOGISTIC REGRESSION

Logistics regression

- Classification problems
  - $y \in \{c_1, c_2, \cdots, c_K\}$
- Binary-classification
  - $y \in \{0,\ 1\}$
- An example
  - Spam filter
  - $\boldsymbol{x}$: → features
  - y: → label
  - $y = 1$: → positive label
  - $y = 0$: → negative label

Logistics regression

- 线性分类器

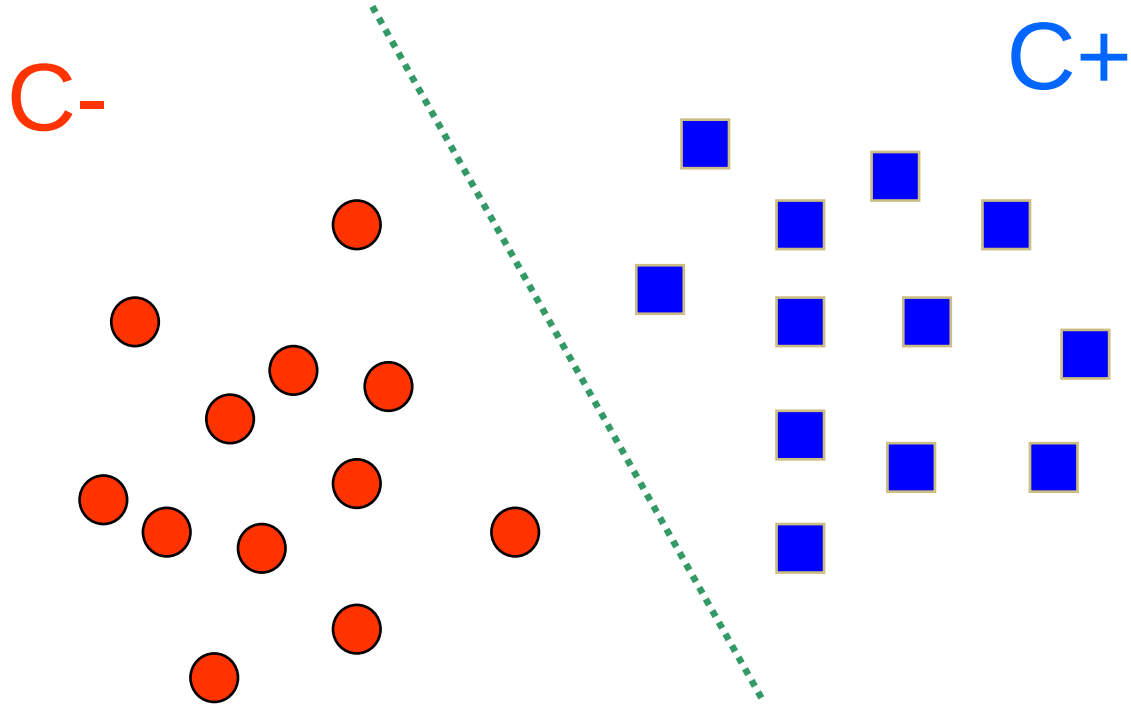


线性回归的目标

$$y=\theta^T x = \theta_0 + \sum_{j=1}^{n} \theta_j x_j$$

Logistics regression

- 线性分类器 $y=\theta^T x = \theta_0 + \sum_{j=1}^{n} \theta_j x_j$



- $y > 0$： → positive label
- $y < 0$： → negative label

Logistics regression

- 问题：能否架起连接线性回归问题和二分类问题的桥梁?
- 最理想的函数—单位阶跃函数

$$y = \begin{cases} 0, & z < 0; \\ 0.5, & z = 0; \\ 1, & z > 0, \end{cases}$$

  - 预测值大于零，判为正例
  - 预测值小于零，判为反例
  - 预测值为零，则无法判别（可任意判别，小概率）
- 单位阶跃函数的缺点
  - 不连续，确定其模型参数θ困难

Logistics regression

- 能否把二分类问题转换为回归问题

  - Spam filter
  - $\boldsymbol{x}$: → features
  - y: → label
  - $y = 1$： → positive label
  - $y = 0$： → negative label

  - Spam filter
  - $\boldsymbol{x}$: → features
  - y: → the probability of (say) positive class
    that is, $y \in [0, 1]$

- 这两类问题有何差异

  原问题：离散　　新问题：连续

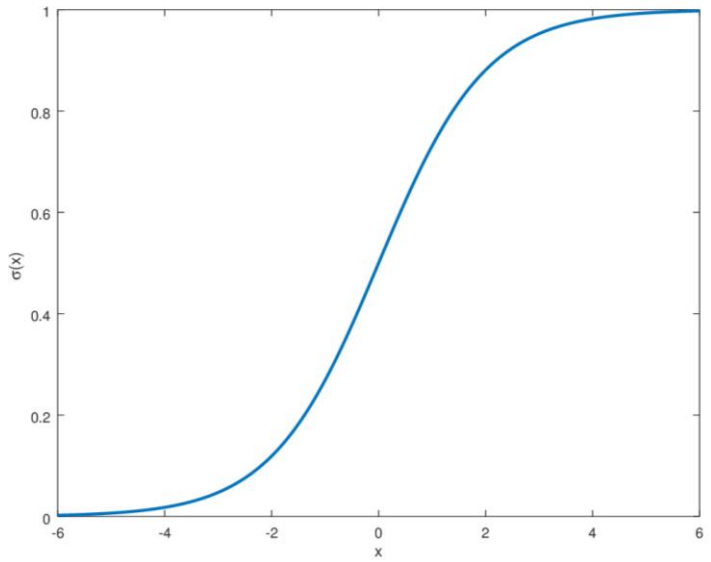
Logistic regression

$$z = \theta^T x = \theta_0 + \sum_{j=1}^{n} \theta_j x_j$$

$$h_\theta(z) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

$$g(z) = \frac{1}{1 + e^{-z}}$$

logistic function or
sigmoid function



Logistic regression

- Properties of logistic function
  - $\lim_{z \to \infty} g(z) = 1$
  - $\lim_{z \to -\infty} g(z) = 0$
  - $0 < g(z) < 1$
  - $g(z)$ with a useful property of the derivative:

$$g(z) = \frac{1}{1 + e^{-z}}$$

$$g'(z) = g(z)(1 - g(z)).$$

Logistic regression

- Properties of logistic function
  - $g(z)$ with a useful property of the derivative:

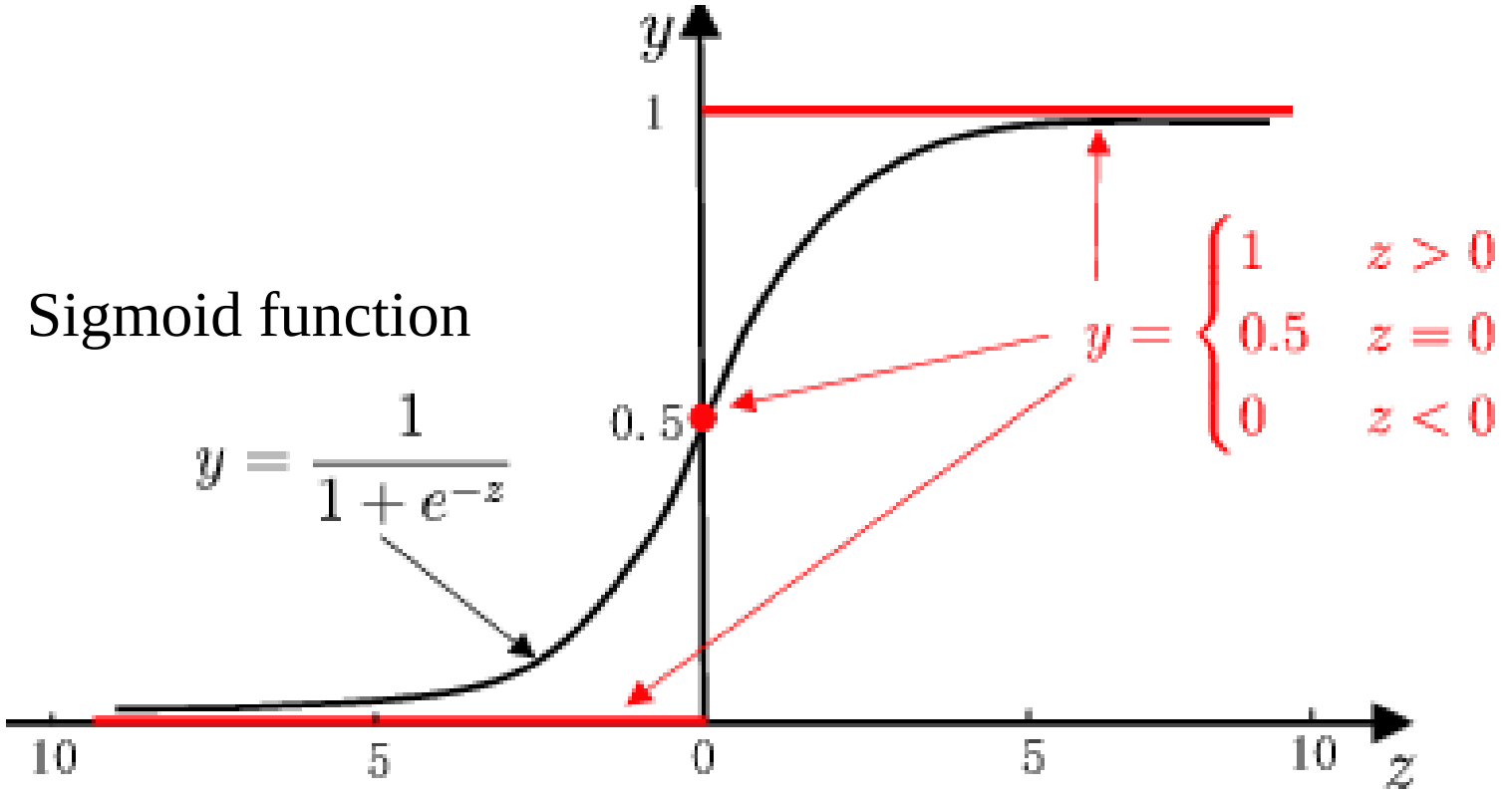
$$g(z) = \frac{1}{1+e^{-z}}$$

$$\begin{aligned} g'(z) &= \frac{d}{dz}\frac{1}{1+e^{-z}} \\ &= \frac{1}{(1+e^{-z})^2}\left(e^{-z}\right) \\ &= \frac{1}{(1+e^{-z})}\cdot\left(1-\frac{1}{(1+e^{-z})}\right) \\ &= g(z)(1-g(z)). \end{aligned}$$

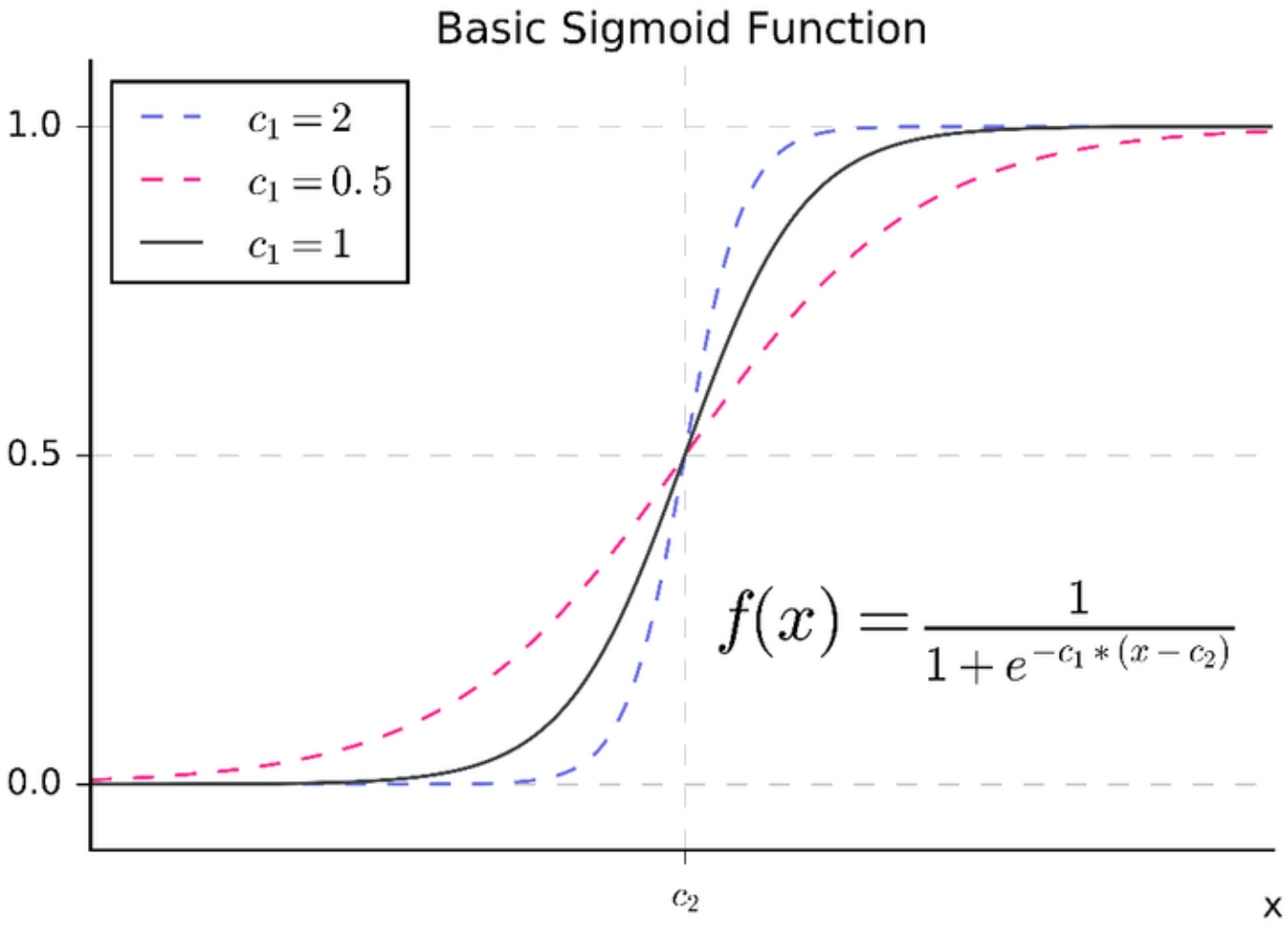
求导过程

Logistic regression

- 单位阶跃函数 vs. Sigmoid function



Logistic regression



Logistic regression

- Probabilistic assumption of binary classification

$$P(y = 1 \mid x; \theta) = h_\theta(x)$$
$$P(y = 0 \mid x; \theta) = 1 - h_\theta(x)$$

- Compact form of binary classification: $y \in \{0, 1\}$

$$p(y \mid x; \theta) = (h_\theta(x))^y \, (1 - h_\theta(x))^{1-y}$$

Logistic regression

- Output with sigmoid function

$$P(y = 1|\boldsymbol{x}; \theta) = h_\theta(\boldsymbol{x}) = \frac{1}{1+e^{-\theta^T x}}$$

$$P(y = 0|\boldsymbol{x}; \theta) = 1 - h_\theta(\boldsymbol{x}) = \frac{e^{-\theta^T \boldsymbol{x}}}{1 + e^{-\theta^T \boldsymbol{x}}}$$

Logistic regression

- Logistic regression

$$P(y = 1|\boldsymbol{x}; \theta) = h_\theta(\boldsymbol{x}) = \frac{1}{1 + e^{-\theta^T \boldsymbol{x}}}$$

- Odds of positive samples

Odds ratio

$$\frac{P(y = 1|\boldsymbol{x}; \theta)}{1 - P(y = 1|\boldsymbol{x}; \theta)} = e^{\theta^T \boldsymbol{x}}$$

- Logit Transformation

Logit Transformation

$$\ln \frac{P(y = 1|\boldsymbol{x}; \theta)}{1 - P(y = 1|\boldsymbol{x}; \theta)} = \ln \frac{h_\theta(\boldsymbol{x})}{1 - h_\theta(\boldsymbol{x})} = \theta^T \boldsymbol{x}$$

Explaining Odds Ratios https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2938757/

Logistic regression

- Log odds(几率比)
  - 几率比是衡量两个事件发生几率的比值，通常用于比较两组之间的差异或关联性。在统计学中，"几率"指的是某事件发生的可能性与不发生的可能性之比。因此，几率比是指在两个不同条件下，某事件发生几率与不发生几率的比值之比。

	事件B发生	事件B未发生
事件A发生	a	b
事件A未发生	c	d

  - 其中，a、b、c 和 d 分别代表不同条件下事件发生的频次。事件A发生时 B事件发生的几率为 a/b ，事件 A 未发生时 B事件发生的几率为 c/d 。于是，几率比OR计算公式为:

$$OR = \frac{a/b}{c/d} = \frac{ad}{bc}$$

Logistic regression

- Log odds
  - 样本作为正例的相对可能性的对数

$$\ln\frac{y}{1-y}=\ln\frac{h_\theta(\boldsymbol{x})}{1-h_\theta(\boldsymbol{x})}=\theta^T\boldsymbol{x}$$

- **例如：p=0.8, success; 1-p=0.2, false，则**

p/(1-p) = 0.8/0.2 = 4,

也可定义： $\mathrm{p}=\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$，$1-\mathrm{p}=\frac{1}{1+e^{\beta_0+\beta_1 x}}$

- **对数几率回归优点**
  - 无需事先假设数据分布
  - 得到“类别”的近似概率预测
  - 可直接应用现有数值优化算法求取最优解

能否直接学习

Logistic regression

- 似然函数

1. 似然与概率的区别

似然 (likehood) 与概率 (probability) 在英语语境中是可以互换的。但是在统计学中，二者有截然不同的用法。

概率描述了已知参数时的随机变量的输出结果；似然则用来描述已知随机变量输出结果时，未知参数的可能取值。

例如，对于“一枚正反对称的硬币上抛十次”这种事件，我们可以问硬币落地时十次都是正面向上的“概率”是多少；而对于“一枚硬币上抛十次”，我们则可以问，这枚硬币正反面对称的“似然”程度是多少。

区别似然和概率的直接方法为，"谁谁谁的概率"中谁谁谁只能是事件，也就是，事件(发生)的概率是多少；而"谁谁谁的似然"中的谁谁谁只能是参数，比如说，参数等于某个值时的似然是多少。

Logistic regression

- 似然函数

2. 似然与概率的联系

先看似然函数的定义。关于参数 θ 的似然函数 （在数值上）等于给定参数θ 后变量 data 的概率（两者的相等并不是说两个函数是同一个，只是数值上的相等）：

$$L(\theta|data) = P(data|\theta) = \prod_{i=1}^{N} P(x_i|\theta)$$

$$data = (x_1, x_2, \ldots, x_n)$$

似然函数的主要用法在于比较它相对取值，虽然这个数值本身不具备任何含义。例如，考虑一组样本，当其输出固定时，这组样本的某个未知参数往往会倾向于等于某个特定值，而不是随便的其他数，此时，似然函数是最大化的。

似然函数乘以一个正的常数之后仍然是似然函数，其取值并不需要满足归一化条件

$$\sum_{x} \alpha \cdot L(\theta|x) \neq 1, \alpha > 0$$

Logistic regression

- 似然函数

3.最大似然估计

最大似然估计是似然函数最初也是最自然的应用。似然函数取得最大值表示相应的参数能够使得统计模型最为合理。从这样一个想法出发，最大似然估计的做法是：首先选取似然函数（一般是概率密度函数或概率质量函数），整理之后求最大值。实际应用中一般会取似然函数的对数作为求最大值的函数，这样求出的最大值和直接求最大值得到的结果是相同的。似然函数的最大值不一定唯一，也不一定存在。

Logistic regression

- Assuming that we have *m* **training examples generated independently**, then

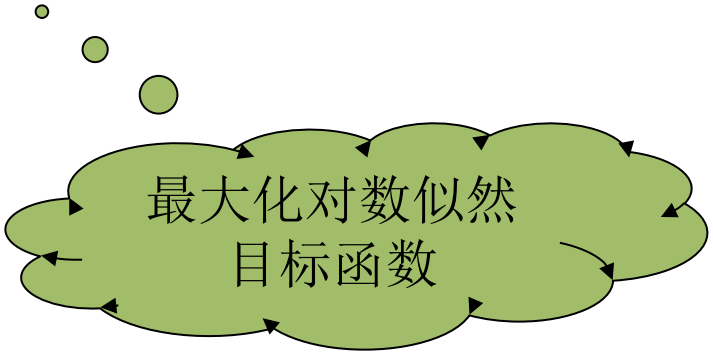
$$\max_{\theta} L(\theta) = p(\vec{y} \mid X; \theta)$$
$$= \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)}; \theta)$$
$$= \prod_{i=1}^{m} \left(h_\theta(x^{(i)})\right)^{y^{(i)}} \left(1 - h_\theta(x^{(i)})\right)^{1-y^{(i)}}$$

最大化似然目标函数

Logistic regression

- It will be easier to maximize the log likelihood:

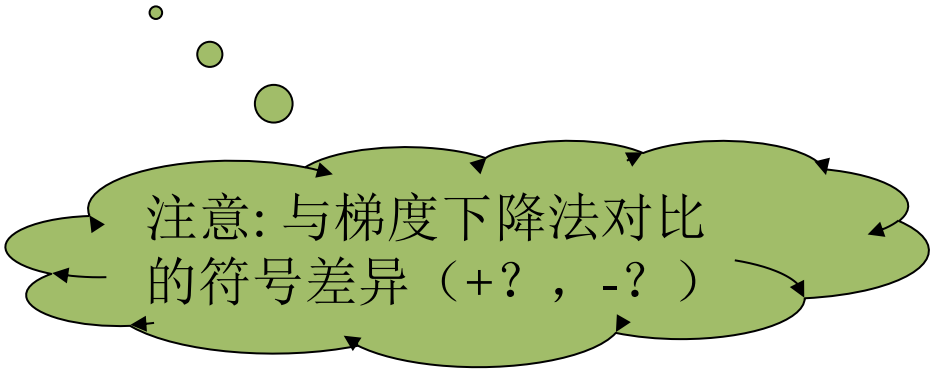
$$\max_{\theta} l(\theta) = \log L(\theta)$$
$$= \sum_{i=1}^{m} y^{(i)} \log h(x^{(i)}) + (1 - y^{(i)}) \log(1 - h(x^{(i)}))$$



Logistic regression

- Gradient ascent

$$\theta := \theta + \alpha \nabla \ell(\theta)$$



Logistic regression

- Again, if only one sample, then

$$
\begin{aligned}
\frac{\partial}{\partial \theta_j}\ell(\theta) &= \left(y\frac{1}{g(\theta^T x)} - (1-y)\frac{1}{1-g(\theta^T x)}\right)\frac{\partial}{\partial \theta_j} g(\theta^T x) \\
&= \left(y\frac{1}{g(\theta^T x)} - (1-y)\frac{1}{1-g(\theta^T x)}\right) g(\theta^T x)(1-g(\theta^T x))\frac{\partial}{\partial \theta_j}\theta^T x \\
&= \left(y(1-g(\theta^T x)) - (1-y)g(\theta^T x)\right) x_j \\
&= \left(y - h_\theta(x)\right) x_j
\end{aligned}
$$

$$\dot{g(z)} = g(z)(1-g(z))$$

Logistic regression

- Stochastic gradient **ascent** for Logistic regression

$$\theta_j := \theta_j + \alpha\left(y^{(i)} - h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right)\right)x_j^{(i)}$$

似曾相识？
殊途同归？

Some comments on LSM

$$h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) = \begin{cases} \boldsymbol{\theta}^T\boldsymbol{x}^{(i)} \\ g\left(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)}\right) = \dfrac{1}{1+e^{-\boldsymbol{\theta}^T\boldsymbol{x}^{(i)}}} \end{cases}$$

Linear regression

Logistic regression

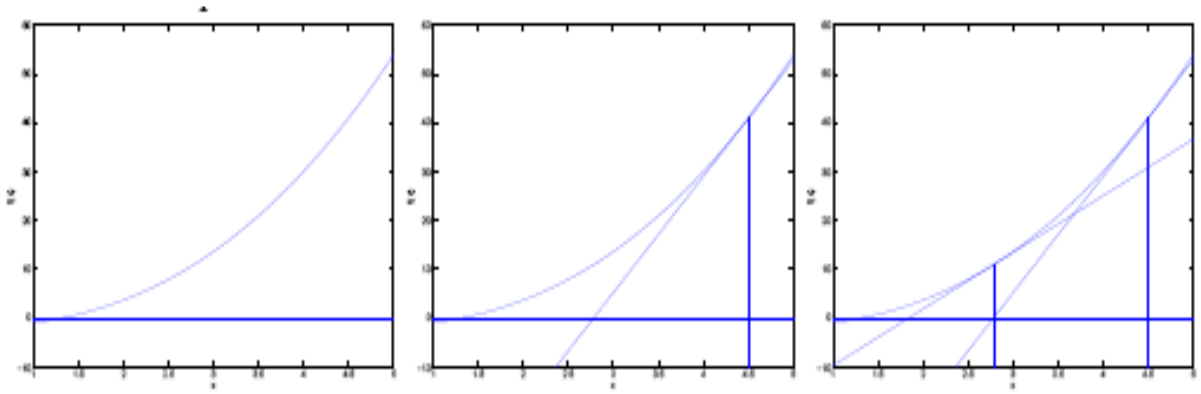
牛顿法
求解对数似然函数的极大值点

Newton's method

- Function of Newton's method
  - Gives a way of getting to $f(\theta) = 0$
- If $\theta \in R$, to get $f(\theta) = 0$, Newton's method performs the following update:

$$\theta := \theta - \frac{f(\theta)}{f'(\theta)}$$

- Here's a picture of the Newton's method in action



Newton's method

- What if we want to use it to maximize some function ℓ
  - let $f(\theta) = l'(\theta) = 0$, then Newton's method is as:

$$\theta := \theta - \frac{l'(\theta)}{l''(\theta)}$$

Newton-Raphson 方法

  - when $\boldsymbol{\theta}$ is a vector

$$\theta := \theta - H^{-1}\nabla_\theta l(\theta)$$

where $\nabla_\theta l(\theta)$ is the vector of partial derivatives of $\ell(\theta)$ with respect to the $\theta_i's$, H called the Hessian, and

$$H_{ij} = \frac{\partial^2 l(\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}$$

牛顿法比批量梯度下降法快，当n不大时，比较有效

- When applied to maximize the logistic regression log likelihood function $\ell(\theta)$, the resulting method is also called **Fisher scoring**.

牛顿法

- 如何求解逻辑回归对数似然函数的极大值点
  - 令 $f(\theta) = l'(\theta) = 0$，则牛顿迭代公式为

$$\theta := \theta - \frac{\ell'(\theta)}{\ell''(\theta)}$$

Newton-Raphson 方法

  - 当θ为向量时

$$\theta := \theta - H^{-1}\nabla_\theta \ell(\theta)$$

其中 $\nabla_\theta \ell(\theta)$是$l(\theta)$偏导数向量，$H$是Hessian矩阵，且

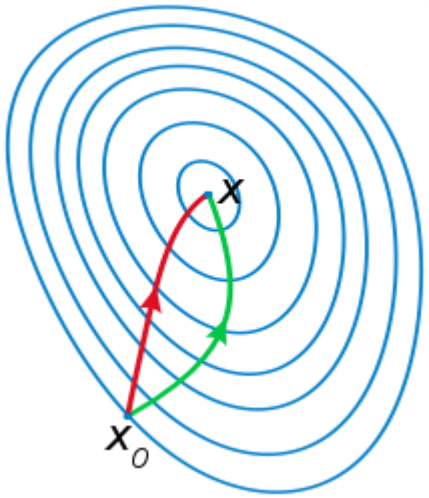
$$H_{ij} = \frac{\partial^2 \ell(\theta)}{\partial\theta_i \partial\theta_j}$$

牛顿法比批量梯度下降法快，当参数量n不大时，比较有效

求解对数似然函数的极大值点

- 牛顿法最大化logistic 回归的log 似然函数: Fisher scoring

$$\theta := \theta - \frac{\ell'(\theta)}{\ell''(\theta)}$$



- When applied to maximize the logistic regression log likelihood function ℓ(θ), the resulting method is also called **Fisher scoring**.

线性回归的概率解释

线性回归的概率解释

- 最小二乘回归为何是非常自然的算法

$$y^{(i)} = \theta^T x^{(i)} + \epsilon^{(i)}$$

$$\varepsilon^{(i)} \sim N(0, \sigma^2)$$

误差项：随机噪声，非模型误差。
误差一般满足均值为0的正态分布

线性回归的概率解释

- 误差项的概率密度

$$p(\epsilon^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\epsilon^{(i)})^2}{2\sigma^2}\right)$$

$$p(y^{(i)}|x^{(i)};\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)$$

$$y^{(i)} \mid x^{(i)};\theta \sim \mathcal{N}(\theta^T x^{(i)}, \sigma^2)$$

线性回归的概率解释

- **似然函数(**likelihood function**)**：

$$
\begin{aligned}
L(\theta) = L(\theta; X, \vec{y}) &= p(\vec{y}|X;\theta) \\
&= \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)};\theta) \\
&= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)
\end{aligned}
$$

- 选择 θ 最大化**似然函数** L(θ)

线性回归的概率解释

- 最大化对数似然函数：$\ell(\theta)$

$$
\begin{aligned}
\ell(\theta) &= \log L(\theta) \\
&= \log \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right) \\
&= \sum_{i=1}^{m} \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right) \\
&= m \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{\sigma^2} \cdot \frac{1}{2} \sum_{i=1}^{m} (y^{(i)} - \theta^T x^{(i)})^2
\end{aligned}
$$

$J(\theta)$

线性回归的概率解释

- 最大化对数似然函数$\ell(\theta)$ ↔最小化 J (θ)

$$\frac{1}{2}\sum_{i=1}^{m}\left(y^{(i)}-\theta^{T}x^{(i)}\right)^{2}$$

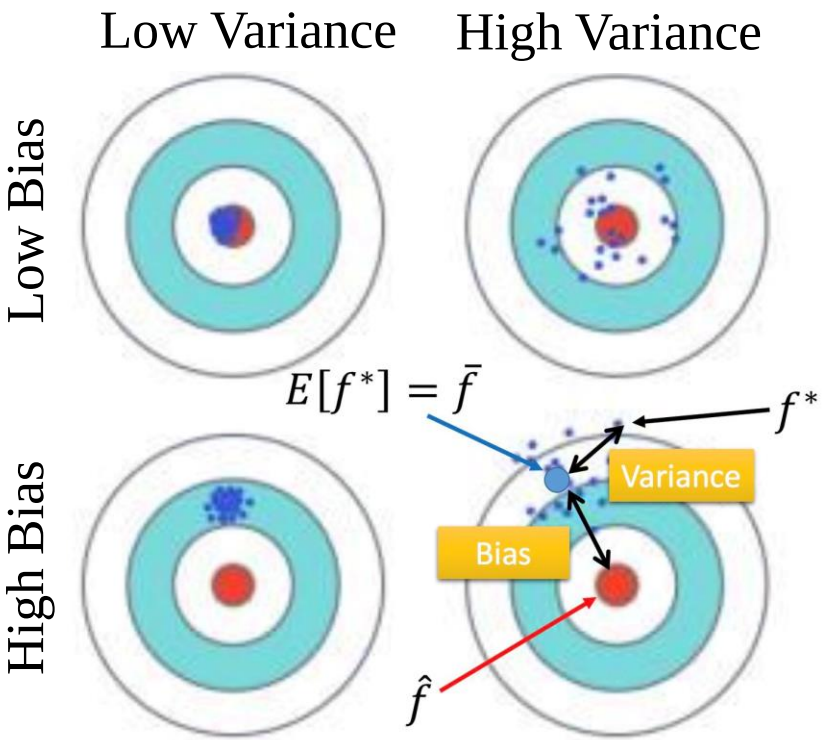
最小二乘回归对应着关于参数θ的最大似然估计

Note however that the probabilistic assumptions are **by no means necessary** for least-squares to **be a perfectly good and rational procedure**, and there may—and indeed there are—other natural assumptions that can also be used to justify it.

模型评估方法和性能评价指标

模型评估方法和性能评价指标

- 为什么
  - 由于偏差过大导致的模型欠拟合以及方差过大导致的过拟合的存在，为了解决这两个问题，需要一整套方法及评价指标。其中评估方法用于评估模型的泛化能力，而性能指标则用于评价单个模型性能的高低。



模型评估方法和性能评价指标

Bias指的是模型在处理训练数据时的错误或误差。具体来说，它衡量了模型的预测结果与实际结果之间的平均差异。一个高偏差的模型往往对训练数据拟合不足，不能捕捉到数据的复杂性和特征。例如，一个线性模型用于拟合非线性的数据分布，就会产生较高的偏差。

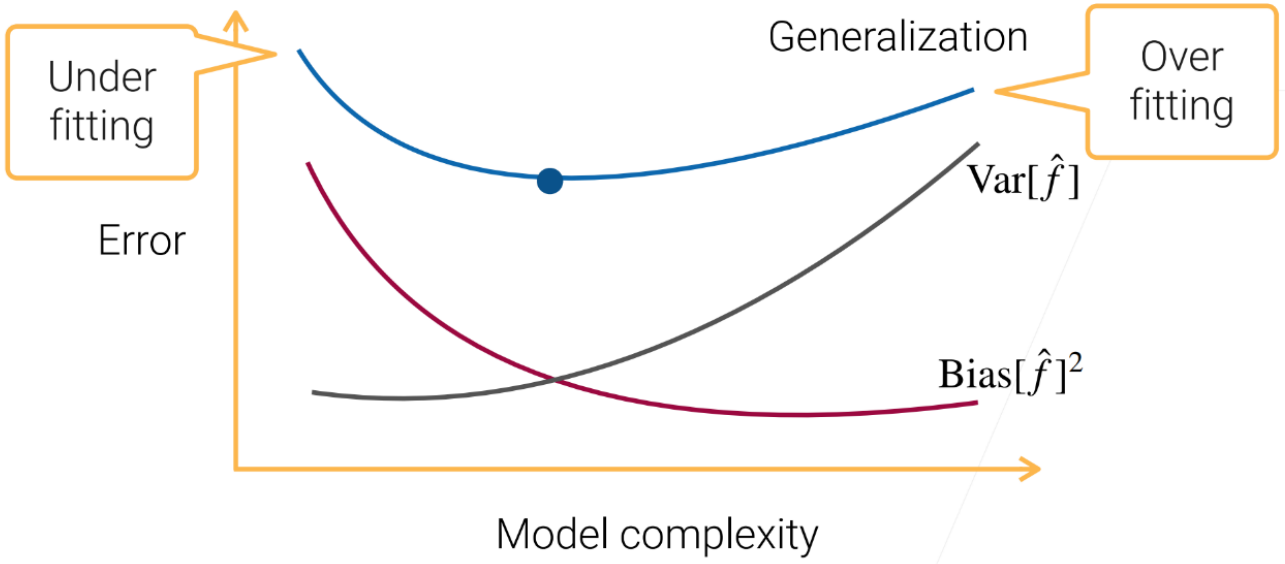
Variance指的是模型对训练数据的敏感性和波动性。它衡量了模型在不同训练数据集上的预测结果的变化程度。一个高方差的模型往往过度拟合了训练数据，对数据中的噪声和随机性过于敏感，导致在新数据上的预测性能较差。

$$E(f; D) = bias^2(x) + var(x) + \varepsilon^2$$

- 期望泛化误差=方差+偏差+噪声
- 偏差刻画学习器的拟合能力
- 方差体现学习器的稳定性
- 噪声表示当前任务期望泛化误差的下界，体现了问题的难度

模型评估方法和性能评价指标

Bias和Variance通常存在一个权衡关系，称为"偏差-方差权衡"。在训练机器学习模型时，我们追求的目标是在偏差和方差之间找到一个平衡点，以实现更好的泛化性能。



模型评估方法—样本集的划分

- 留出法
  - 分层采样
  - 2/3~4/5 training
  - >30 test
- 交叉验证 k-fold cross validation
  - K个互斥的子集
  - 分层采样
  - K-1 training
  - 1 test
  - K=10
  - K=样本数： 留一法 leave-one-out

模型评估方法

- 自助法 bootstrapping
  - 可重复采样 or 有放回的采样
  - 大约有1/3的测试集
  - 适用数据集规模较小
  - 对集成学习有利
  - 会引入估计偏差

性能评价指标

二分类效果评估方法

- 准确率（accuracy）
- 精确率-查准率（precision）
- 召回率-查全率（recall）
- 综合评价指标（F1 measure）
- ROC，AUC值（Receiver Operating Characteristic ROC，Area Under Curve，AUC）

真阳性（true positives），真阴性（true negatives），假阳性（false positives），假阴性（false negatives）。阳性和阴性指分类，真和假指预测的正确与否。

Confusion matrix (contingency table)

TP = true positives(真阳性)*:* If an object is positive and it is classified as positive

FN = false negatives(真阴性)*:* If an object is positive and it is classified as negative

TN = true negatives(假阳性)*:* If the object is negative and it is classified as negative

FP = false positives(假阴性)*:* If the object is negative and If it is classified as positive

性能评价指标

- 准确率 Accuracy：所有分类正确的样本占全部样本的比例

$$Accuracy = (TP+TN)/(TP+TN+FP+FN)$$

- 精准率又叫做：Precision、查准率:预测的正例中有多少对的

$$Precision = TP/(TP+FP)$$

- 召回率Recall (真阳性率、查全率、敏感性)：真正的正例中有多少对的

$$Recall = TP/(TP+FN)$$

性能评价指标

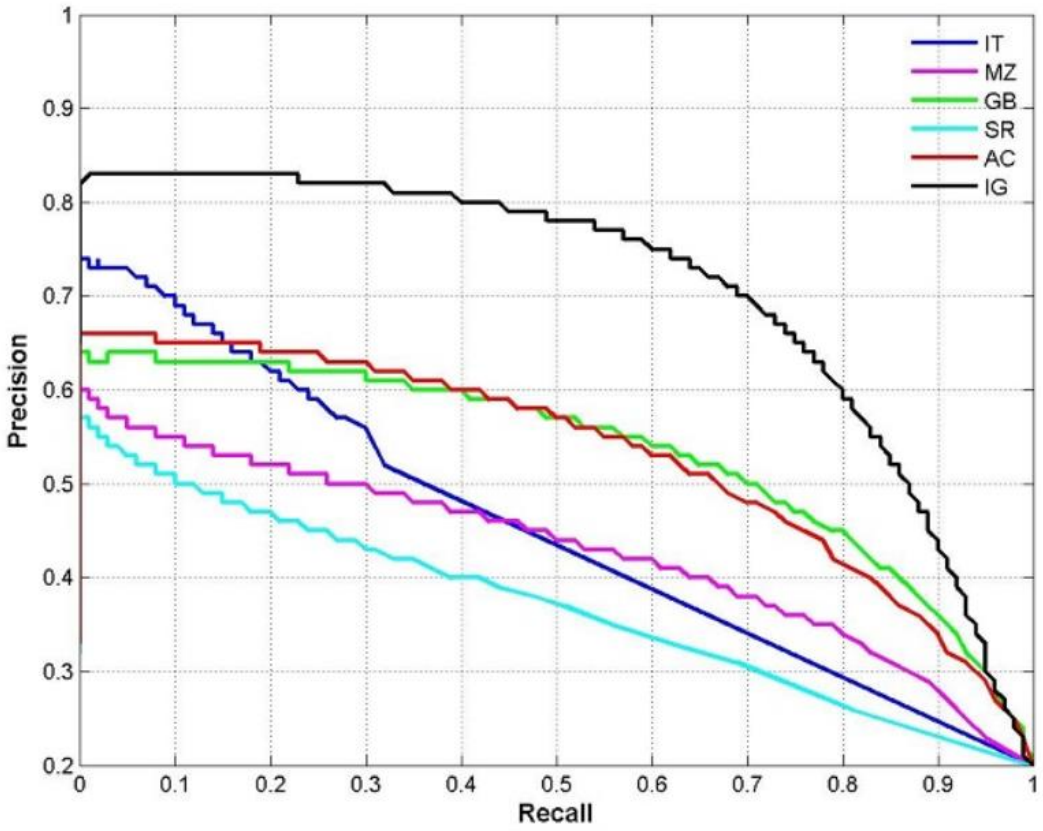
- 不同问题对于 P 与 R 的偏重程度不一样；
- 典型例子：
  - 推荐系统——担心用户不胜其烦，结果尽可能是用户真正感兴趣的

    查准率 P　重要
  - 重大疾病筛查——不可遗漏

    查全率 R　重要

性能评价指标

- P-R curve：根据预测结果将预测样本排序，最有可能为正样本的在前，最不可能的在后，依次将样本预测为正样本，分别计算当前的精确率和召回率，绘制P-R曲线。



性能评价指标

- F1 度量：P与R的调和平均

$$\frac{1}{F1}=\frac{1}{2}\left(\frac{1}{P}+\frac{1}{R}\right)$$

$$F1=\frac{2*P*R}{P+R}=\frac{2*TP}{样本总数+TP-TN}$$

性能评价指标

- 敏感性 Sensitivity(真阳性率，召回率或查全率)：True-positive rate (TPR) **(真正例率)**

$$TPR = TP/(TP+FN)$$

- 特异性 Specificity (真阴性率)：

TN/(TN+FP)= true negatives/actual negatives

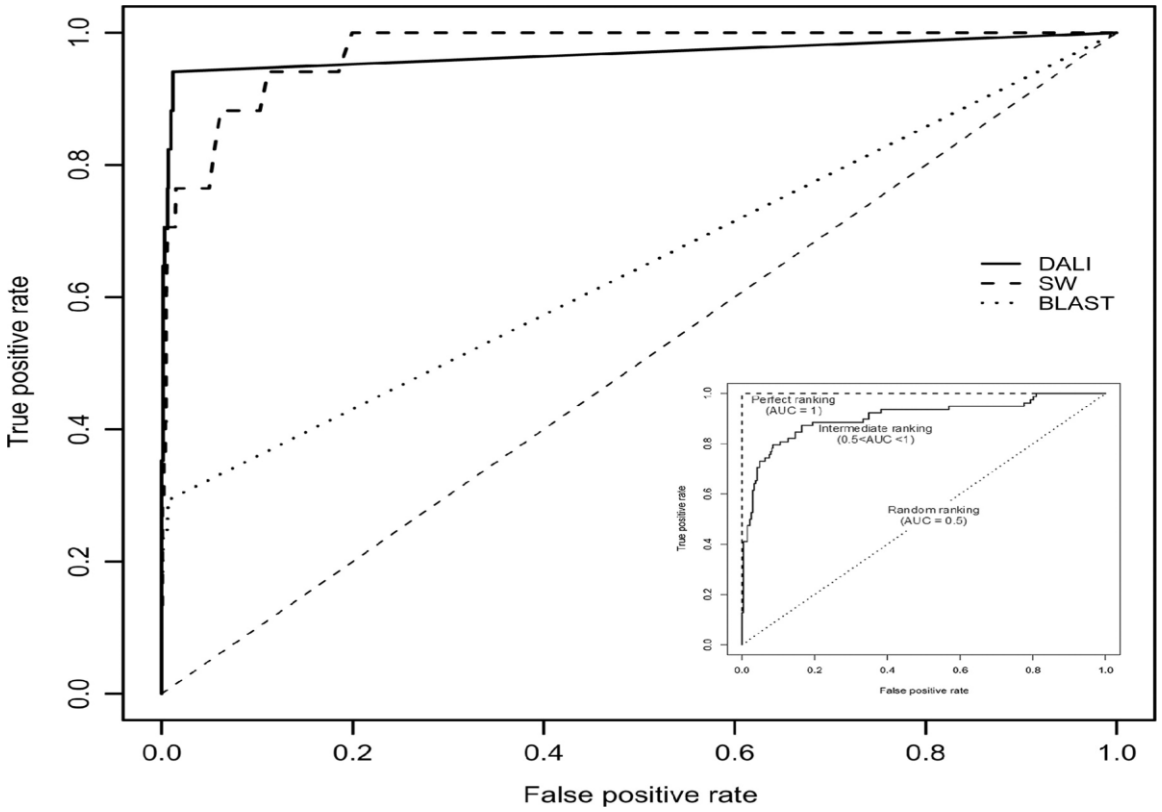
- 假正例率 False positive rate (FPR)

$$FPR = FP/(TP+FP)$$

TN maybe Large ！Non-informative

性能评价指标

- **receiver operating characteristic curve, ROC（area under curve, AUC）**



Sonego P et al. Brief Bioinform 2008;9:198-209

性能评价指标

- **(Matthews) Correlation coefficient:**

$$CC = \frac{(TP)(TN) - (FP)(FN)}{\sqrt{(PP)(PN(AP)(AN)}}$$

AP = actual positives=TP+FN

AN = actual negatives=TP+TN

PP = predicted positives=TP+FP

PN = predicted negatives=FN+TN

1.3 MCC 的特性:

范围： MCC 的值范围在 -1 到 1 之间，其中 1 表示完美预测，0 表示等同于随机预测，-1 表示完全相反的预测。

适用性： MCC 尤其适用于样本不平衡的二分类问题，因为它同时考虑了各个类别的预测性能。

全面性： 与其他指标相比，MCC 更全面地考虑了分类模型在各个方面的性能，对于综合评估模型的质量提供了更全面的视角。

作业

下载垃圾邮件数据集（http://archive.ics.uci.edu/ml/datasets/Spambase）。请从spambase.csv读入数据。数据集基本信息如下：样本数: 4601,特征数量: 57,类别：1为垃圾邮件，0为非垃圾邮件。将样本集划分为70%的训练集，30%作为测试集。使用Matlab或Python库中提供函数分别完成以下内容：

使用最大似然估计优化方法求解模型，分别算出准确率（accuracy）、精确率-查准率（precision）、召回率-查全率（recall）、综合评价指标（F1 measure），并画出 P-R曲线和AUC曲线。