机器学习基础

代启国

大连民族大学
计算机科学与技术系

2018 年 10 月

机器学习简介

机器学习 (Machine Learning, ML)

- 是一门多领域交叉学科，涉及计算机、高等数学、概率与统计、线性代数等多门学科。

机器学习简介

机器学习 (Machine Learning, ML)

- 是一门多领域交叉学科，涉及计算机、高等数学、概率与统计、线性代数等多门学科。
- 专门研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能。

机器学习简介

机器学习 (Machine Learning, ML)

- 是一门多领域交叉学科，涉及计算机、高等数学、概率与统计、线性代数等多门学科。
- 专门研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能。
- 简单来讲，就是计算机从数据中学习出规律和模式，以应用在新数据上做预测的任务。

机器学习简介

机器学习 (Machine Learning, ML)

- 是一门多领域交叉学科，涉及计算机、高等数学、概率与统计、线性代数等多门学科。
- 专门研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能。
- 简单来讲，就是计算机从数据中学习出规律和模式，以应用在新数据上做预测的任务。
- 是目前人工智能的主要方向，是使计算机具有智能的主要途径，其应用遍及人工智能的各个领域。

机器学习主要问题

- 有监督学习（Supervised）
  - 分类（Classification）
  - 回归（Regression）
- 无监督学习（Unsupervised）
  - 聚类（Clustering）
- 半监督学习（Semi-Supervised）
- 增强学习（Reinforcement Learning）
- 其它

本课程主要关注的问题

- 有监督学习（Supervised）
  - 分类（Classification）
  - 回归（Regression）
- 无监督学习（Unsupervised）
  - 聚类（Clustering）

监督学习

监督学习（supervised learning）

- 给定已标记的训练数据

$$D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_m, y_m)\}$$

其中，每个训练样本 i 由它的特征（通常为向量）$\mathbf{x_i}$ 和一个期望的输出值（也称为监督信号）y_i 组成。

监督学习

监督学习（supervised learning）

- 给定已标记的训练数据

$$D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_m, y_m)\}$$

其中，每个训练样本 i 由它的特征（通常为向量）$\mathbf{x_i}$ 和一个期望的输出值（也称为监督信号）y_i 组成。

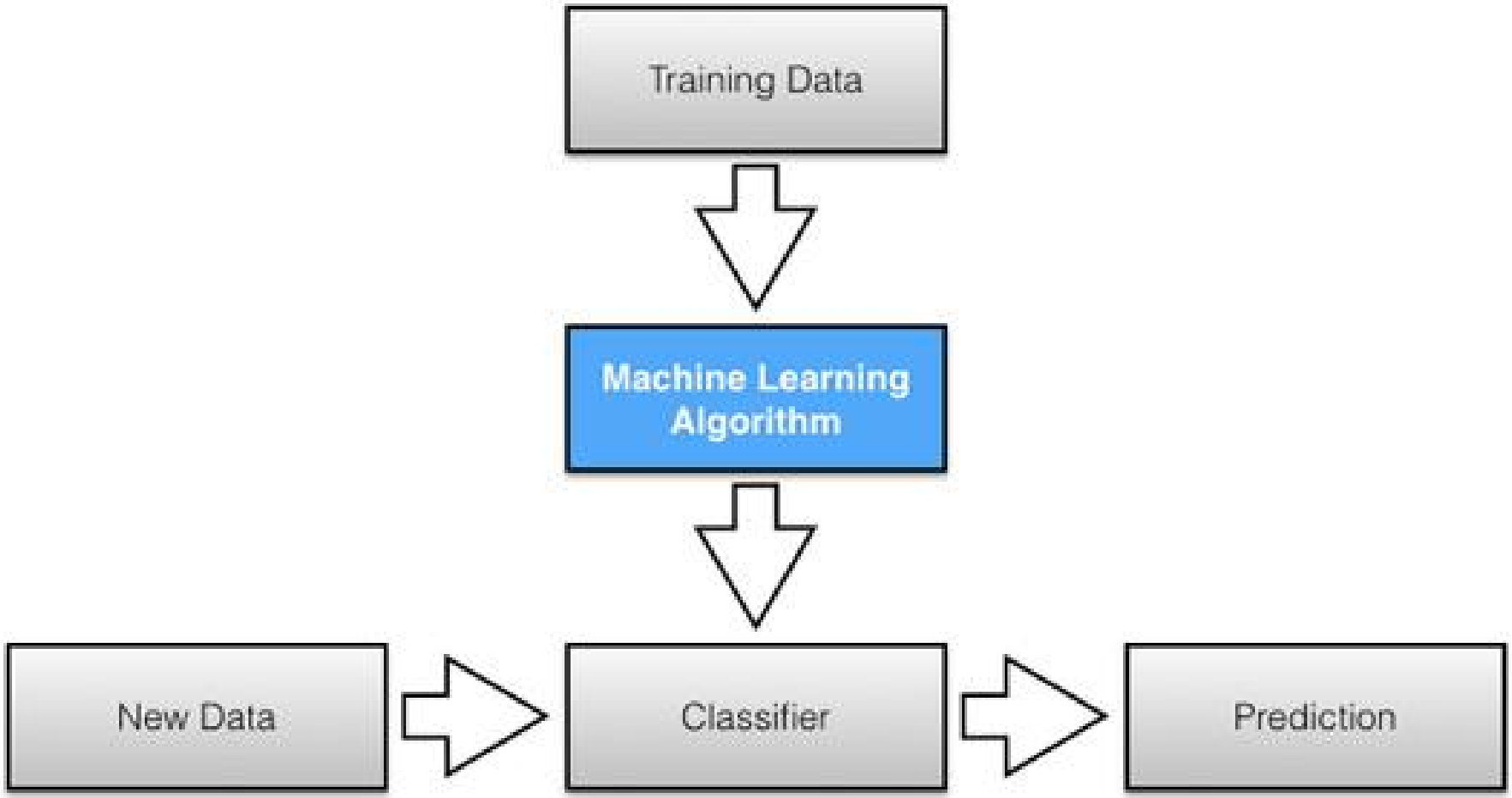
- 从 D 中训练模型

$$y = f(\mathbf{x})$$

，来推断一个新样本的输出值的机器学习任务。

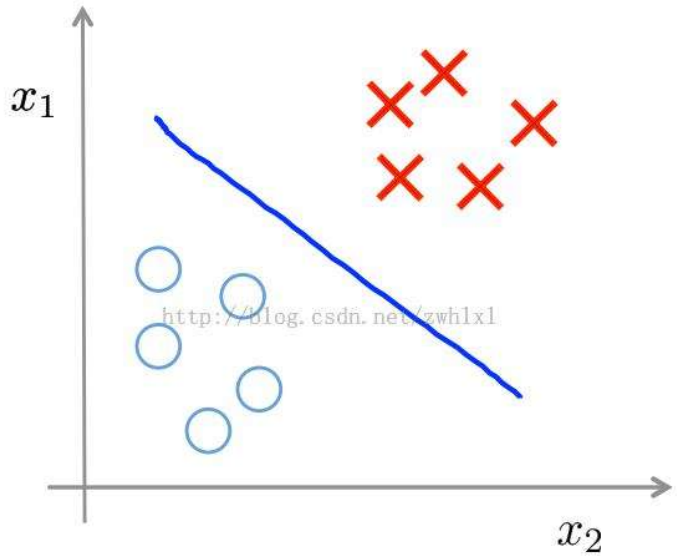
监督学习

监督学习（supervised learning）



监督学习（supervised learning）

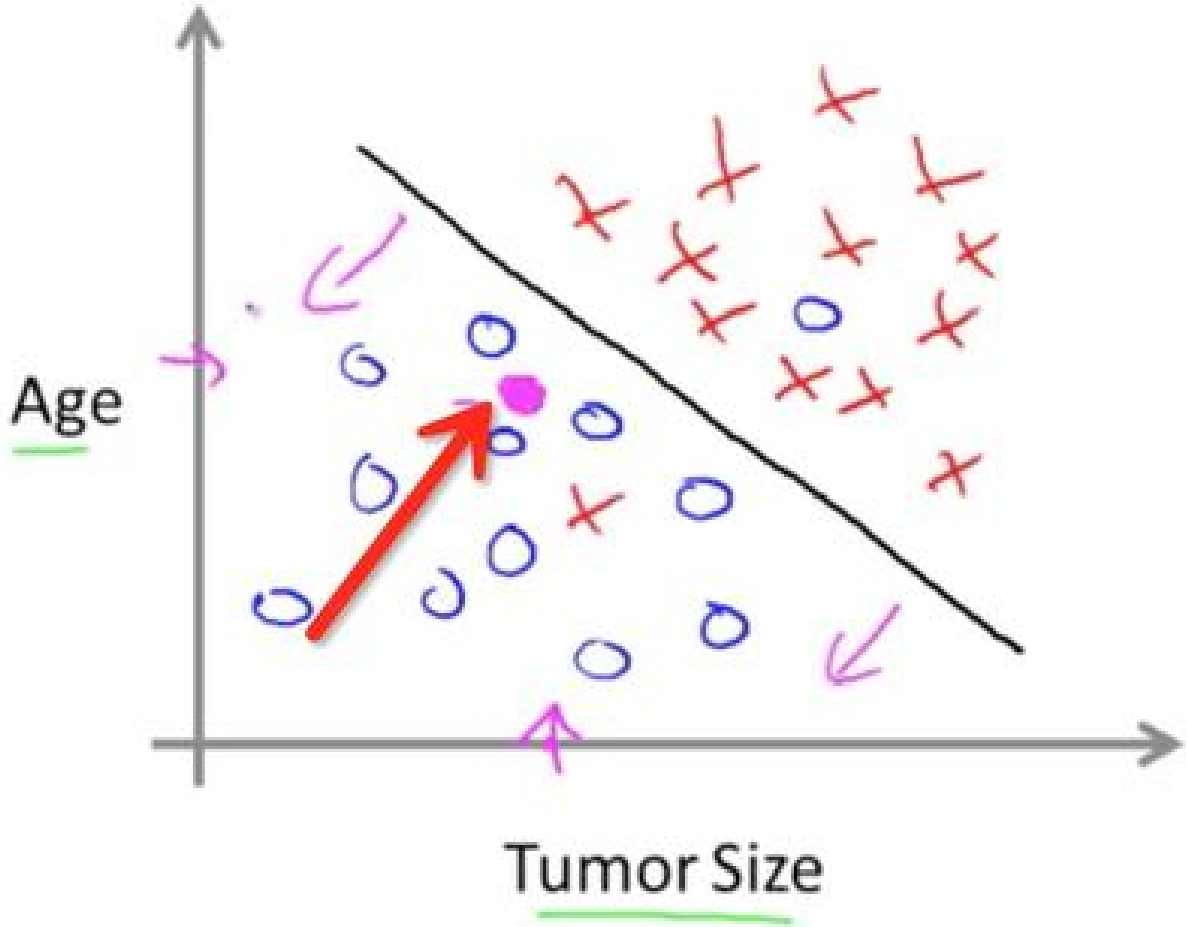
Supervised learning



Training set: $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), (x^{(3)}, y^{(3)}), \ldots, (x^{(m)}, y^{(m)})\}$

监督学习

监督学习（supervised learning）



模型评估与选择

- 评价一个机器学习模型（有监督）的好坏需要特定的评估方法，并据此对模型进行选择，从而得到一个更好的模型。
- 误差
  - 经验误差 (Training error) 或训练误差（Training error）
  - 泛化误差（Generalization error）
- 拟合（Fitting）
  - 过拟合（Overfitting）
  - 欠拟合（Underfitting）
- 评估方法（Validation）
- 性能度量（Performance measure）

误差（Error）

训练误差 (training error) 或经验误差 (empirical error)

- 学习器的实际预测输出与训练样本的真实输出之间的误差

泛化误差 (generalization error)

- 学习器的实际预测输出与新样本的真实输出之间的误差

学习的目标

- 我们希望得到泛化误差小的学习器
- But, 我们事先并不知道新样本是什么样
- 我们能做的————努力使经验误差最小化
- 经验最小化会导致一定“风险”

拟合问题

- 过拟合 (Overfitting)
  - 学习器把训练样本学得“太好”了，导致在新样本上泛化性能下降 (预测准确率降低)

拟合问题

- 过拟合 (Overfitting)
  - 学习器把训练样本学得“太好”了，导致在新样本上泛化性能下降 (预测准确率降低)
- 欠拟合 (Underfitting)
  - 学习器在训练样本上学得“太差”了，同样导致在新样本上泛化能力下降

拟合问题

- 过拟合 (Overfitting)
  - 学习器把训练样本学得“太好”了，导致在新样本上泛化性能下降 (预测准确率降低)
- 欠拟合 (Underfitting)
  - 学习器在训练样本上学得“太差”了，同样导致在新样本上泛化能力下降

过拟合与欠拟合的直观对比

拟合问题

- 欠拟合 (Underfitting)
  - 学习能力过于低下
  - 通过增加训练的迭代次数等方式提高学习能力

拟合问题

- 欠拟合 (Underfitting)
  - 学习能力过于低下
  - 通过增加训练的迭代次数等方式提高学习能力
- 过拟合 (Overfitting)
  - 学习能力过于强大
  - 机器学习领域面临的关键障碍
  - 无法彻底避免，只能“缓解”（减小经验风险）

拟合问题

- 欠拟合 (Underfitting)
  - 学习能力过于低下
  - 通过增加训练的迭代次数等方式提高学习能力
- 过拟合 (Overfitting)
  - 学习能力过于强大
  - 机器学习领域面临的关键障碍
  - 无法彻底避免，只能“缓解”（减小经验风险）

模型选择（Model selection）

- 机器学习中有很多算法，甚至一种算法在使用不同参数时也会产生不同 (学习器) 模型
- 最优方案是选择使得泛化误差最小的模型进行应用
- 矛盾如何解决？现实训练机器学习时如何选择不同模型?

评估方法

为了对模型（学习器）的泛化误差进行评估：

- 测试数据集（Testing set）
  - 测试模型对测试数据集的判别能力

评估方法

为了对模型（学习器）的泛化误差进行评估：

- 测试数据集（Testing set）
  - 测试模型对测试数据集的判别能力
  - 以测试集上的“测试误差”（testing error）作为模型的近似“泛化误差”

评估方法

为了对模型（学习器）的泛化误差进行评估：

- 测试数据集（Testing set）
  - 测试模型对测试数据集的判别能力
  - 以测试集上的“测试误差”（testing error）作为模型的近似“泛化误差”

假设

- 测试集中样本时从真实样本中独立同分布采样而得
- 测试集应该与训练集互斥，即：测试样本为在训练过程中被使用

评估方法

为了对模型（学习器）的泛化误差进行评估：

- 测试数据集（Testing set）
  - 测试模型对测试数据集的判别能力
  - 以测试集上的“测试误差”（testing error）作为模型的近似“泛化误差”

假设

- 测试集中样本时从真实样本中独立同分布采样而得
- 测试集应该与训练集互斥，即：测试样本为在训练过程中被使用

思考

- 如果老师出了 10 道题给学生训练，老师又用这 10 道题作为考试题
- 考试成绩能否反映出学生们学习得好不好?

评估方法

为了对模型（学习器）的泛化误差进行评估：

- 测试数据集（Testing set）
  - 测试模型对测试数据集的判别能力
  - 以测试集上的“测试误差”（testing error）作为模型的近似“泛化误差”

假设

- 测试集中样本时从真实样本中独立同分布采样而得
- 测试集应该与训练集互斥，即：测试样本为在训练过程中被使用

思考

- 如果老师出了 10 道题给学生训练，老师又用这 10 道题作为考试题
- 考试成绩能否反映出学生们学习得好不好?
- 答案是否定的。因为我们需要”举一反三”的学习能力。

评估方法

为了对模型（学习器）的泛化误差进行评估：

- 测试数据集（Testing set）
  - 测试模型对测试数据集的判别能力
  - 以测试集上的“测试误差”（testing error）作为模型的近似“泛化误差”

如何利用现有数据构建测试集?

- 现有数据集: $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_m, y_m)\}$
- 从现有数据集中取出一部分作为测试集 T
- 其余部分作为训练集 S
- 训练集与测试集互斥：$S \cup T = D$ 并且 $S \cap T = \emptyset$

评估方法

常见的模型评估方法：

- 留出法（Hold-out）
- 交叉验证法 (Cross validation)

留出法

- 将数据集 D 划分成训练集 S 和测试集 T：$\boldsymbol{S} \cup \boldsymbol{T} = \boldsymbol{D}$ 并且 $\boldsymbol{S} \cap \boldsymbol{T} = \emptyset$
- 以二分类为例，假定 D 包含 1000 个样本，S 中包含 700 个样本，T 中包含 300 样本
- 在 S 上训练的模型在 T 上有 90 个样本分类错误，错误率是 30%
- 测试集与训练集要保持数据一致性
  - 正负样本比例要一致
  - 如果 D 中样本是有序的，。。。
- 单次留出法评估不可靠，通常需要多次随机划分，重复实验后取平均

评估方法

常见的模型评估方法：

- 留出法（Hold-out）
- 交叉验证法 (Cross validation)

交叉验证法

- k-折交叉验证（k-fold cross validation）
- 把数据平均分成互斥的 k 份
- 通常 $k = 10$;
- 如果 $k = m$（m 为 D 中的样本数），称为“留一法”

评估方法

参数

- 大多数机器学习算法都有一些参数（Parameter）需要设定
- 参数配置不同，所学习模型的性能往往有显著差别

评估方法

参数

- 大多数机器学习算法都有一些参数（Parameter）需要设定
- 参数配置不同，所学习模型的性能往往有显著差别

参数类别

- 算法参数（超参数）：通常需要人工设定
- 模型参数：通过算法学习

评估方法

参数

- 大多数机器学习算法都有一些参数（Parameter）需要设定
- 参数配置不同，所学习模型的性能往往有显著差别

参数类别

- 算法参数（超参数）：通常需要人工设定
- 模型参数：通过算法学习

参数调节（Parameter tuning）

- 学习算法有很多超参数是实数值，遍历所有参数是不可能

评估方法

参数

- 大多数机器学习算法都有一些参数（Parameter）需要设定
- 参数配置不同，所学习模型的性能往往有显著差别

参数类别

- 算法参数（超参数）：通常需要人工设定
- 模型参数：通过算法学习

参数调节（Parameter tuning）

- 学习算法有很多超参数是实数值，遍历所有参数是不可能
- 常用方法：选定一定范围和步长
  - 例：[0，0.2] 范围，以 0.05 为步长，需要评估的模型超参数是 5 个

评估方法

参数

- 大多数机器学习算法都有一些参数（Parameter）需要设定
- 参数配置不同，所学习模型的性能往往有显著差别

参数类别

- 算法参数（超参数）：通常需要人工设定
- 模型参数：通过算法学习

参数调节（Parameter tuning）

- 学习算法有很多超参数是实数值，遍历所有参数是不可能
- 常用方法：选定一定范围和步长
  - 例：[0，0.2] 范围，以 0.05 为步长，需要评估的模型超参数是 5 个
- 训练一个理想的机器学习模型，调参需要很大的工作量

性能度量

性能度量（Performance measure）

- 除上述评估方法外，还需要衡量模型泛化能力的评价标准

回归问题 的性能度量

性能度量

性能度量（Performance measure）

- 除上述评估方法外，还需要衡量模型泛化能力的评价标准

回归问题 的性能度量

- 均方误差（Mean squared error）

$$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}(f(x_i)-y_i)^2$$

性能度量

性能度量（Performance measure）

- 除上述评估方法外，还需要衡量模型泛化能力的评价标准

回归问题 的性能度量

- 均方误差（Mean squared error）

$$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}\left(f(x_i)-y_i\right)^2$$

- 其描述的一般形式

$$E(f;D)=\int_{\mathbf{x}\in D}\left(f(\mathbf{x})-y\right)^2 p\left(\mathbf{x}\right)d\mathbf{x}$$

性能度量

分类问题 性能度量的常用方法

- 错误率与精度
- 查准率、查全率与 $F1$
- ROC 与 AUC

性能度量

分类问题性能度量的常用方法

- 错误率与精度
  - 既适用于二类分类，也适用于多类分类任务
  - 错误率：分类错误的样本数占总样本数的比例

$$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\left(f(\mathbf{x}_i)\neq y_i\right)$$

  - 精度：分类正确的样本数占总样本数的比例

$$acc(f;D)=\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\left(f(\mathbf{x}_i)=y_i\right)$$

性能度量

分类问题性能度量的常用方法

- 错误率与精度
  - 既适用于二类分类，也适用于多类分类任务
  - 错误率：分类错误的样本数占总样本数的比例

$$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\left(f(\mathbf{x}_i)\neq y_i\right)$$

  - 精度：分类正确的样本数占总样本数的比例

$$acc(f;D)=\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\left(f(\mathbf{x}_i)=y_i\right)$$

错误率与精度之关系

$$E(f;D)=1-acc(f;D)$$

性能度量

- 查准率、查全率与 $F1$
  - 查准率（Precision，P）

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

  - 查全率（Recall，R）

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

性能度量

- 查准率、查全率与 $F1$
  - 查准率（Precision，P）

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

  - 查全率（Recall，R）

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

表: 分类结果混淆矩阵

真实情况	预测结果	
	正例	反例
正例	TP（真正例）	FN（假反例）
反例	FP（假正例）	TN（真反例）

性能度量

- 查准率、查全率与 $F1$
  - 查准率（Precision，P）

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

  - 查全率（Recall，R）

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

表：分类结果混淆矩阵

真实情况	预测结果	
	正例	反例
正例	TP（真正例）	FN（假反例）
反例	FP（假正例）	TN（真反例）

查全率与查准率之矛盾

以搜索引擎为例，

- 查准率：检索出的信息中有多少比例是用户感兴趣的；
- 查全率：用户感兴趣的信息中有多少被检索出来；

性能度量

- 查准率、查全率与 $F1$
  - 查准率（Precision，P）

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

  - 查全率（Recall，R）

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

  - F1：P 和 R 的调和平均数

$$F1 = \frac{2 \times P \times R}{P + R}$$

表：分类结果混淆矩阵

真实情况	预测结果	
	正例	反例
正例	TP（真正例）	FN（假反例）
反例	FP（假正例）	TN（真反例）

查全率与查准率之矛盾

以搜索引擎为例，

- 查准率：检索出的信息中有多少比例是用户感兴趣的；
- 查全率：用户感兴趣的信息中有多少被检索出来；

性能度量

ROC 与 AUC

- 很多分类器的预测值是概率值，该值大于一定阈值（threshold）为正类，否则为反类
  - 例，神经网络一般情况下对每个样本预测出一个 $[0.0, 1.0]$ 之间的实值，预测值与阈值（$t = 0.5$）进行比较，大于 t 为正类，小于 t 为反类
- 按照预测的概率将样本进行降序排序
  - 最有可能是正例的排在最前面
- 分类问题：在该序列中以某截断点（cut point）将样本分为两部分
  - 前一部分被预测为正例，后一部分为反例
- 对于不同分类任务需求，截断点位置不同
- 需要综合考虑不同任务下的泛化性能好坏

性能度量

ROC 与 AUC

- 受试者工作特征（Receiver Operating Characteristic，ROC）

性能度量

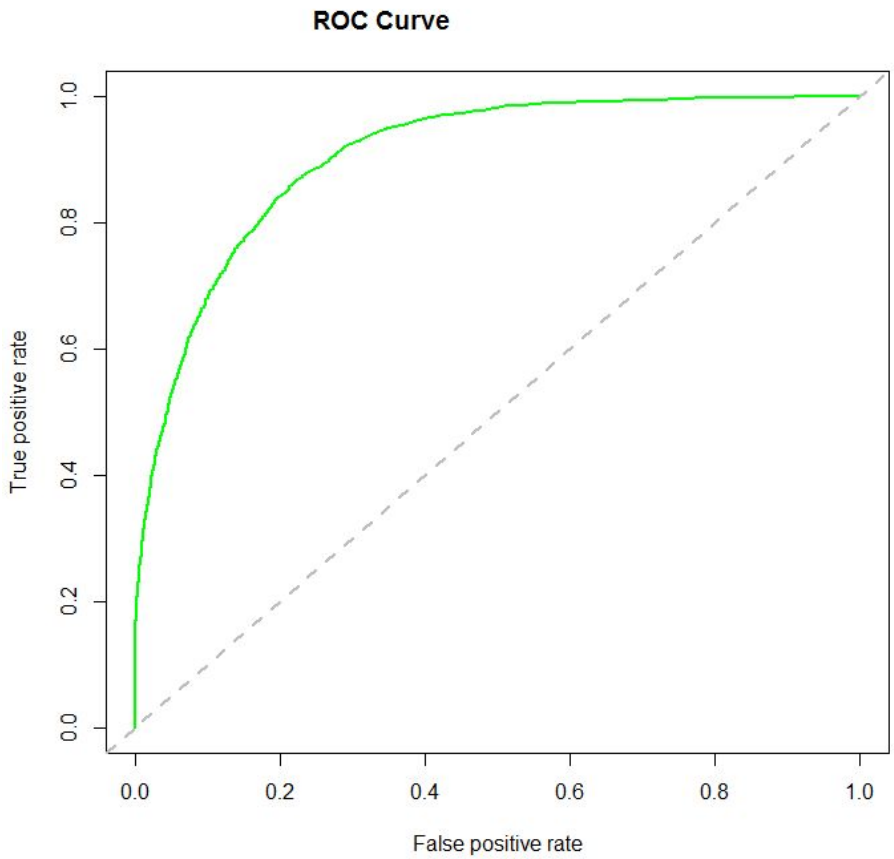
ROC 与 AUC

- 受试者工作特征（Receiver Operating Characteristic，ROC）
- 根据预测概率对样本进行排序，依次逐个把样本作为正例，计算两个量，得出 ROC 曲线
  - 真正例比率（TPR）

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

  - 假正率比率（FPR）

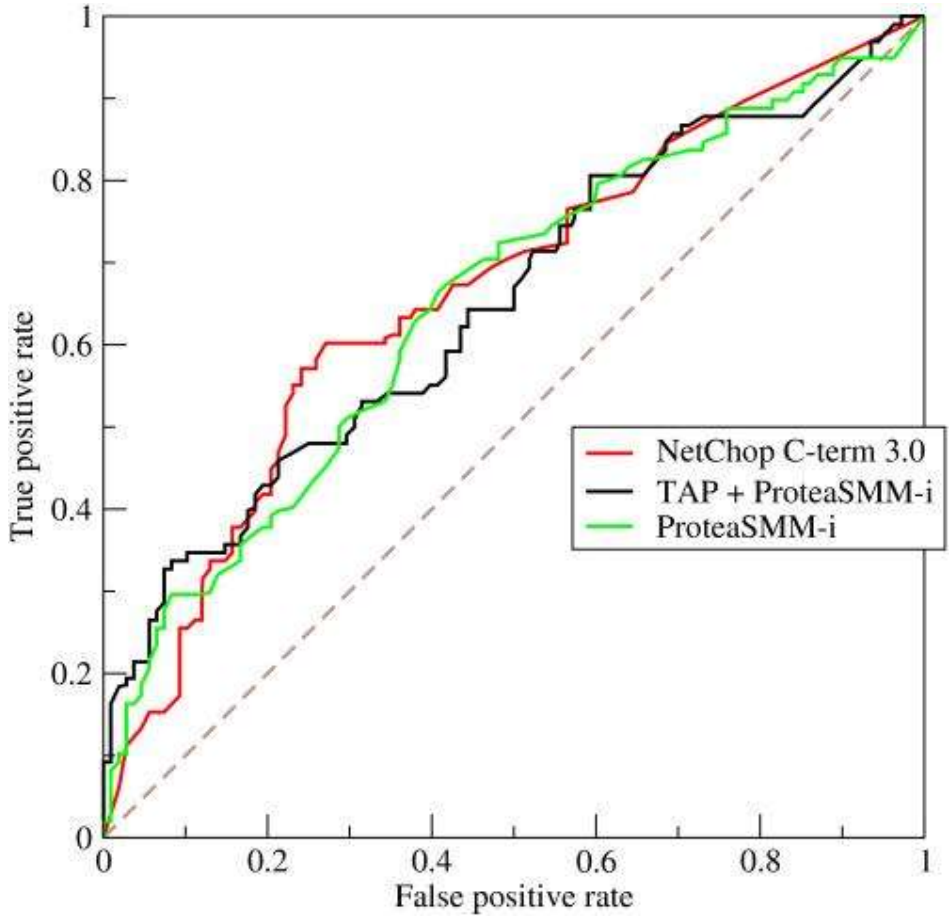
$$FPR = \frac{FP}{TN + FP}$$



性能度量

ROC 与 AUC

- 对角线（虚线）：随机猜想模型
- 点（0,1）：将所有正例排在所有反例之前的“理想模型”



性能度量

ROC 与 AUC

- 对角线（虚线）：随机猜想模型
- 点（0,1）：将所有正例排在所有反例之前的“理想模型”

比较不同学习模型的 ROC 曲线

- 如果一个学习器的曲线完全“包住”另外一个，则可断言其性能由于后者；
- 如果两条线交叉，一般需要计算“线下面积”，即 AUC (Area Under ROC Curve)