分类模型

授课教师：吴翔
邮箱：wuhsiang@hust.edu.cn

March 18-25, 2019

统计学习概述

统计学习方法

统计机器学习（statistical machine learning）可分为：

- 有监督学习（supervised learning）vs 无监督学习（unsupervised learning）：聚类分析即为典型的无监督学习
- 参数方法（parametric methods）vs 非参数方法（non-parametric methods）
- 回归（regression）问题 vs 分类（classification）问题：分别针对连续变量和分类变量

测试均方误差的分解

测试均方误差的期望值（expected test MSE）可以分解为如下三个部分：

$$E(y-\hat{f}(x))^2 = \underbrace{\mathrm{Var}(\hat{f}(x))}_{\text{variance}} + \underbrace{[\mathrm{Bias}(\hat{f}(x))]^2}_{\text{bias}} + \underbrace{\mathrm{Var}(\epsilon)}_{\text{irreducible}} .$$

- 模型方差（variance）：针对不同的训练数据，$\hat{f}$ 的变化程度。
- 模型偏误（bias）：通过相对简化的模型来**近似**真实世界的问题时所引入的误差。

权衡模型偏误与方差

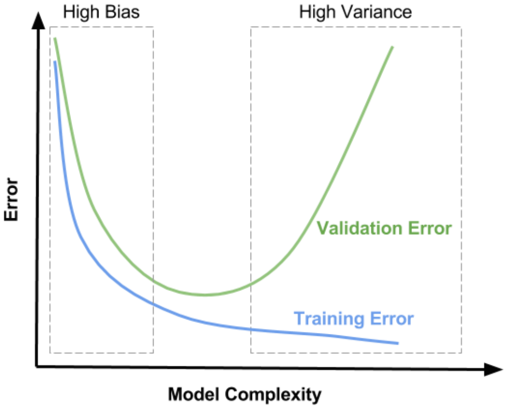

图 1: bias-variance trade-off

如何选择统计模型?

- 传统统计模型的局限：线性回归模型等统计模型通常最小化训练数据的均方误差，但是其测试均方误差（test MSE）却较大。换言之，传统统计模型执着于寻求“真实规律”，以致于将一些随机因素**误判**为 f 的真实性质。
- 权衡模型偏误与方差（bias-variance trade-off）：随着模型灵活性（或自由度）的增加，模型方差随之增大，但模型偏误则相应减小（过度拟合问题）。通过交叉验证（cross-validation）来实现两者的权衡。
- 权衡预测精度与可解释性（accuracy-interpretability trade-off）：诸如 bagging、boosting、support vector machines 等非线性模型具有很高的预测精度，但不易解释；linear models 等易于解释，但预测精度不高。两者的权衡取决于研究目的。

交叉验证

交叉验证将原始数据集分为训练集（training set）和验证集（validation set），并以验证集的错误率选择最佳模型。

- 留一交叉验证法（leave-one-out cross validation, LOOCV）
- k 折交叉验证法（$k-$fold CV）：将观测集随机分为 k 个大小基本一致的组，或说折（fold）。每次选取其中一折作为验证集，而剩余 $k-1$ 折作为训练集。通常，取 $k=5$ 或 $k=10$。

分类模型验证集错误率：

$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\mathrm{Err}_k = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{m_k}\sum_{i=1}^{m_k} I(y_i \neq \hat{y}_i).$$

分类模型概述

预测分类响应变量（categorical response variable）：

1. 基本分类模型（basic classifier）
2. 树模型（tree-based models）
3. 支持向量机（support vector machine, SVM）
4. 聚类模型（clustering models）

分类模型的评价

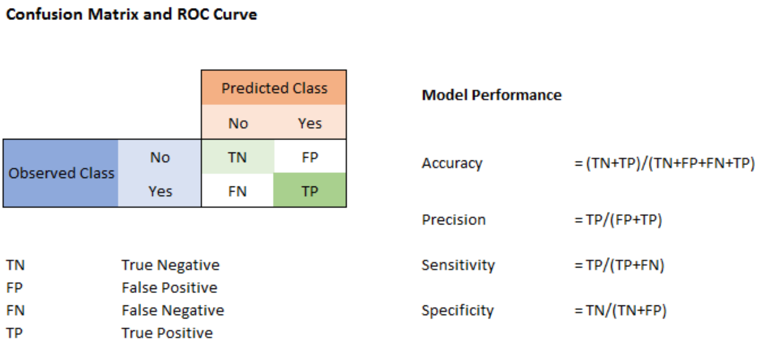

图 2: confusion matrix

混淆矩阵及评价指标

- 准确率（accuracy）：指被正确判断的样本的比例。
- 精确率（precision）：指判断为阳性的样本中，实际为阳性的比例。
- 灵敏度（sensitivity）：也称为真阳性率、召回率（recall rate）。指实际为阳性的样本中，被正确判断为阳性的比例。
- 特异度（specificity）：也称为真阴性率。指实际为阴性的样本中，被正确判断为阴性的比例。

完美的分类器可以达到 100% 的灵敏度，及 100% 的特异度。但是理论上所有的分类器都会有最小的误差范围，称为贝叶斯错误率。

可以绘制接收者操作特征曲线（receiver operating characteristic curve，ROC 曲线），并结合 ROC 曲线下面积（area under the curve of ROC, AUC）来比较不同分类模型。

ROC 曲线与 AUC

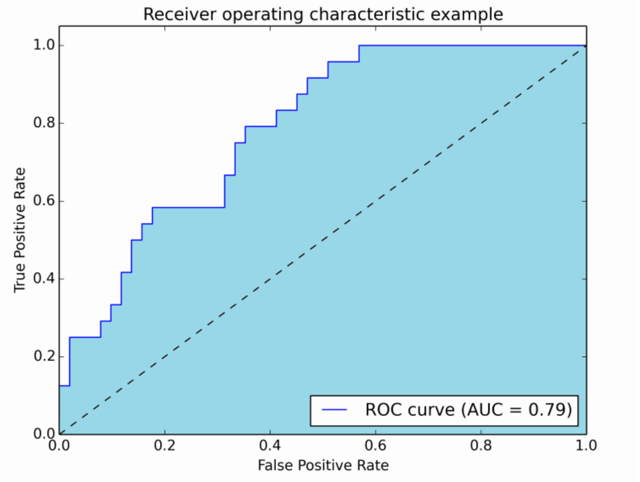

图 3: ROC curve and AUC

基本分类模型

基本分类模型（basic classifier）

1. 逻辑斯蒂回归（logistic regression）
2. 贝叶斯分类器（bayes classifier）
3. 线性判别分析（linear discriminant analysis, LDA）
4. 二次判别分析（quadratic discriminant analysis, QDA）
5. K 最近邻（$K-$nearest neighbor, KNN）

logistic 回归

给定 X 条件下事件 Y 发生的概率 $p(X) = \Pr(Y = 1|X)$，据此可以将发生比（odd）的对数建模为 X 的线性函数

$$\log[\frac{p(X)}{1 - p(X)}] = \beta X.$$

上式左侧称为对数发生比（log-odd）或分对数（logit），其取值范围在 $(-\infty, \infty)$。

当类别 $K \geq 2$ 时，则采用多类别 logistic 回归模型。

似然函数

可以通过**最大似然估计**（maximum likelihood estimation, MLE）得到 logistic 回归的参数值。

参数记为 θ，数据记为 D。似然函数（likelihood function）是参数 θ 的函数，且定义为给定参数 θ 时，观测到数据 D 的概率：

$$l(\theta) = p(D|\theta).$$

例如，logistic 回归模型的似然函数

$$l(\beta) = \prod_{i=1}^{n} p(X_i)^{y_i}[1 - p(X_i)]^{1-y_i}.$$

贝叶斯定理

贝叶斯定理阐述了随机变量 X 和 Y 的条件概率之间的关系：

$$p(Y|X) = \frac{p(X, Y)}{p(X)} = \frac{p(Y) \cdot p(X|Y)}{p(X)}.$$

或从"数据-参数"的视角而言，参数 θ 的后验分布 $\pi(\theta) = p(\theta|D)$ 正比于参数的先验分布 $p(\theta)$ 和似然函数 $l(\theta)$ 之积：

$$\pi(\theta) = \frac{p(\theta)p(D|\theta)}{p(D)} = \frac{p(\theta)l(\theta)}{p(D)}.$$

课堂板书：贝叶斯定理推导及概念解释

贝叶斯定理与分类

对于分类（categorical）响应变量 Y 而言，运用贝叶斯定理：

$$p(Y=k|X=x)=\frac{p(Y=k)\cdot p(X=x|Y=k)}{p(X=x)}.$$

假定 x 是 m 维向量（即特征数量），简写为

$$p(C_k|x)=\frac{p(C_k)\cdot p(x|C_k)}{p(x)}\propto p(C_k)\prod_{i=1}^{m}p(x_i|C_k)$$

贝叶斯分类器

贝叶斯分类器（bayesian classifier）选择后验概率 $p(C_k|x)$ 最大的类别，作为分类结果，即 argmax $p(C_k|x)$。

可以证明，贝叶斯分类器将产生最低的测试错误率，亦即**贝叶斯错误率**。相应用于分类的边界，成为贝叶斯决策边界（bayes decision boundary）。

问题在于，如何推导出后验概率 $p(C_k|x)$? 我们需要更多**假设**。

LDA

线性判别分析（linear discriminant analysis，LDA）假定 $p(x|C_k) \sim N(\mu_k, \Sigma)$。LDA 即是条件概率 $p(x|C_k)$ 为（多元）正态分布时的贝叶斯分类器，其判别函数 $f(x)$ 为线性函数。

考虑 x 是一维的情况，

$$p(x|C_k) = \frac{1}{\sqrt{2\pi}\sigma}\exp[-\frac{1}{2\sigma^2}(x-\mu_k)^2],$$

由此根据后验概率 $p(C_k|x)$ 的对数，得到如下判别函数

$$f_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log[p(C_k)].$$

课堂板书：推导判别函数

LDA 示意图

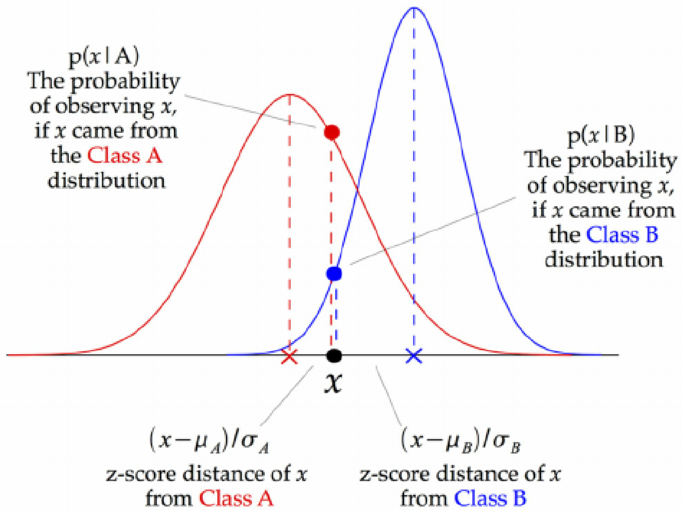

图 4: Illustration of LDA

QDA

二次判别分析（quadratic discriminant analysis, QDA）假定 $p(x|C_k) \sim N(\mu_k, \Sigma_k)$。QDA 即是条件概率 $p(x|C_k)$ 为（多元）正态分布时的贝叶斯分类器，其判别函数 $f(x)$ 为二次函数。QDA 与 LDA 的差别在于，协方差矩阵 Σ_k 是否假定相等。

x 为多维向量时，LDA 的判别函数为

$$f_k(x) = x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log[p(C_k)].$$

相应地，QDA 的判别函数为

$$f_k(x) = -\frac{1}{2}x^T\Sigma_k^{-1}x + x^T\Sigma_k^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma_k^{-1}\mu_k + \log[p(C_k)].$$

QDA 示意图

- 左图：对于两个类别，均有 $\rho(X_1, X_2) = 0.7$
- 右图：对于橙色类别，$\rho(X_1, X_2) = 0.7$；对于蓝色类别，$\rho(X_1, X_2) = -0.7$

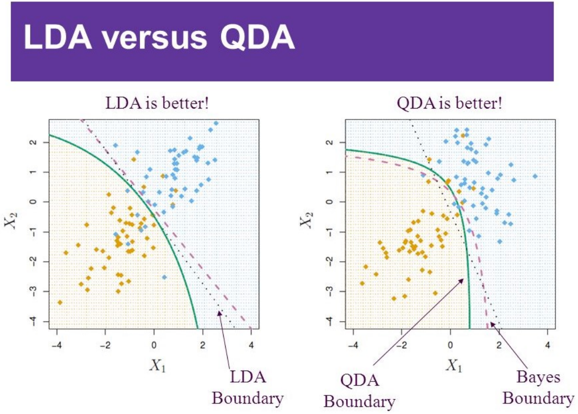

图 5: Illustration of QDA

KNN

通常难以知道 $p(C_k|X)$ 的分布。因而，可以设法估计条件分布 $p(C_k|X)$。

对给定正整数 K 和测试观测值 x_0，K 最近邻（KNN）分类器首先识别训练集中 K 个最靠近 x_0 的点集 A，继而以集合 A 中的点估计条件概率：

$$p(C_j|x_0) == \frac{1}{K}\sum_{i\in A} I(y_i = j).$$

最后，运用贝叶斯规则将测试观测值 x_0 分到后验概率 $p(C_j|x_0)$ 最大的类中。

- KNN 作为典型的非参数方法（non-parametric methods），能够产生一个近似于最优贝叶斯分类器的效果
- K 值的选择对 KNN 分类器的效果有根本影响

KNN 示意图

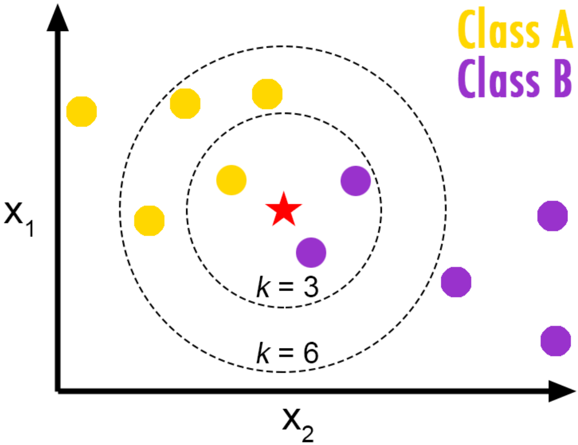

图 6: Illustration of KNN

模型讨论

- LDA 中的判别函数与 logistic 回归中的对数发生比（log-odd）均是 x 的线性函数，因而二者都产生一个线性决策边界，且分类结果相近。若 $p(x|C_k) \sim N(\mu_k, \Sigma)$ 近似成立，则 LDA 优于 logistic 回归；反之，logistic 回归优于 LDA。
- KNN 作为非参数方法，对决策边界的形状没有做出任何假设。因而当决策边界高度非线性时，KNN 优于 LDA 和 logistic 回归。
- QDA 是线性决策边界（LDA 和 logistic 回归）和非参数 KNN 方法的折衷方案，采用了二次函数形式的决策边界。

案例：股票市场走势预测

聚类模型

聚类模型（clustering models）

1. K 均值聚类（$K-$means clustering）
2. 系统聚类（hierarchical clustering）

K 均值聚类

树模型

树模型（tree-based models）

1. 决策树
2. 装袋法（bagging）
3. 随机森林（random forest）
4. 提升法（boosting）

决策树

支持向量机

支持向量机