模式识别与统计学习

本页PDF

以下笔记为个人理解，涵盖了这门课的学习重点，若和教材有出入请以教材为主

从EM算法开始属于无监督学习

统计学习概论

常见的定义

统计学习的定义：是关于计算机基于数据构建概率统计模型并运用模型对数据进行预测和分析

统计学习的对象：数据 - 关于数据的基本假设是**同类数据具有一定的统计规律性**

统计学习的目的：对数据进行预测和分析

统计学习的三要素：方法=模型+策略+算法

期望风险：模型在数据联合分布上的期望损失，衡量模型泛化能力，但实际应用难以求解，使用结构风险代替

经验风险：模型在已知训练集的平均损失

结构风险：经验风险 + 正则化项

泛化误差：衡量模型的泛化能力，等于近似误差 + 估计误差 - 近似误差：模型简单引起的误差，可以看成偏差 - 估计误差：数据太少引起的误差，可以看成方差

欠拟合：高偏差低方差，模型在训练集，验证集，测试集的泛化能力都很差 - 解决办法：增加模型复杂度，增加数据，减少正则化项等

过拟合：低偏差高方差，模型在已知的数据集上预测的很好，但在未知的数据集上预测的很差 - 解决办法：降低模型复杂度，增加数据，增加正则化项，早停等

生成式模型：学习的是数据的联合分布

判别式模型：学习的是数据的条件概率分布或者决策函数

概率模型：学习的是条件概率分布或者联合分布 - 朴素贝叶斯，高斯混合模型

非概率模型：学习的是决策函数 - 有感知机，knn，kmeans，SVM等

注意：逻辑斯特回归既是概率模型也是非概率模型

参数化模型：对数据有基本假设，用带参函数建模，参数量不变 - 有感知机，朴素贝叶斯，逻辑斯特回归等

非参数化模型：对数据没有基本假设，无学习参数 - 有SVM，knn，决策树，Adaboost等

统计学习可以分为：监督学习、无监督学习、强化学习 - 监督学习：从有标注数据中学习预测模型的机器学习问题 - 无监督学习：从无标注数据中学习预测模型的机器学习问题

混淆矩阵

	P	N
T	TP	FP
F	FN	TN

P和N代表样本真实标签，T和F代表样本的预测标签 - TP：T表示样本的预测结果是正确的，P表示样本被预测为正例

要会计算的指标 精确率

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

召回率

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

F1值

$$\frac{2}{F1} = \frac{1}{Precision} + \frac{1}{Recall}$$

感知机

学习的是R^n空间的超平面方程

$$w \cdot x + b = 0$$

其中,$w = [w^{(1)}, w^{(2)}, \dots, w^{(n)}], x = [x^{(1)}, x^{(2)}, \dots, x^{(n)}]^T$

点到超平面的几何距离

$$d = \frac{|w \cdot x + b|}{||w||}$$

$$\begin{cases} w \cdot x + b > 0 & \text{在正面} \\ w \cdot x + b = 0 & \text{在超平面上} \\ w \cdot x + b < 0 & \text{在反面} \end{cases}$$

感知机其实就是线性二分类模型（判别式模型）

$$f(x) = sign(w \cdot x + b)$$

$$sign(x) = \begin{cases} 1 & x \geq 0 \\ -1 & x < 0 \end{cases}$$

损失函数

- 只考虑了误分类点，M是误分类点的集合
- $||w||$不被考虑，能正确分类即可，至于真正距离超平面多远，模型并不关心

$$L(w, b) = -\sum_{x_i \in M} y_i (w \cdot x_i + b)$$

随机梯度下降：每次只使用一个样本进行训练，更新模型参数

$$\min_{w,b} L(w,b) = -\sum_{x_i \in M} y_i(w \cdot x_i + b)$$

梯度

$$\nabla_w L(w,b) = -\sum_{x_i \in M} y_i x_i \quad \nabla_b L(w,b) = -\sum_{x_i \in M} y_i$$

若$y_i(w \cdot x_i + b) \leq 0$，则对参数进行更新，其中$\eta$为学习率

$$w \leftarrow w + \eta y_i x_i \quad b \leftarrow b + \eta y_i$$

对偶形式

将w和b表示为实例x_i和y_i的线性组合的形式，求解其系数

$$w = w_0 + \sum_{i=1}^{N} \alpha_i y_i x_i \quad b = b_0 + \sum_{i=1}^{N} \alpha_i y_i$$

其中：$\alpha_i = n_i \eta$，n_i为x_i被误分类的次数

Gram矩阵 使用Gram矩阵存储内积可以加速计算

$$G = [x_i \cdot x_j]_{N \times N} = \begin{bmatrix} x_1 \cdot x_1 & x_1 \cdot x_2 & \cdots & x_1 \cdot x_N \\ x_2 \cdot x_1 & x_2 \cdot x_2 & \cdots & x_2 \cdot x_N \\ \vdots & \vdots & \ddots & \vdots \\ x_N \cdot x_1 & x_N \cdot x_2 & \cdots & x_N \cdot x_N \end{bmatrix}$$

感知机模型

$$f(x) = sign(\sum_{j=1}^{N} \alpha_j y_j x_j \cdot x + b)$$

其中：$\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_N)^T$

如果$y_i(\sum_{i=1}^{N} \alpha_i y_i x_i \cdot x + b) \leq 0$，进行参数更新

$$\alpha_j \leftarrow \alpha_j + \eta \quad b \leftarrow b + \eta y_i$$

我个人的理解：

如果样本x_i被错误分类了，对α_i进行更新，同时使用更新的α进行下次样本判别

knn

多分类回归模型，时间和空间复杂度高，分类边界的特点是**不规则曲线**

决策函数：决定了样本被分成什么类别

- 当$y_i = c_j$时，$I(y_i = c_j)$为1，否则为0
- $N_k(x)$为数据集中涵盖与x最近k个点的领域
- 统计这k个样本的类别，选取个数最多的类别作为预测结果

$$y = argmax \sum_{x_i \in N_k(x)} I(y_i = c_j)$$

距离度量

- L_1为曼哈顿距离
- L_2为欧氏距离
- L_∞为切比雪夫距离

$$L_k(x,y) = (\sum_{i=1}^{d} |x_i - y_j|^k)^{\frac{1}{k}}$$

k值的选取对模型的影响

- k值太小：模型对噪声敏感，整体模型复杂，易过拟合，近似误差减小，估计误差增大
- k值太大：整体模型简单，易欠拟合，估计误差减小，近似误差增大

当$k = 1$时模型的训练误差是0，当$k > 1$的时候不一定

朴素贝叶斯

贝叶斯定理

$$\begin{aligned} p(x|y) &= \frac{p(y|x)p(x)}{p(y)} \\ p(x|y,z) &= \frac{p(x|z)p(y|x,z)}{p(y,z)} \\ &= \frac{p(x,y,z)}{p(y,z)} \\ &= \frac{p(z)p(x|z)p(y|x,z)}{p(z)p(y|z)} \end{aligned}$$

基本思想

- 已知类条件概率密度参数表达式和先验概率
- 利用贝叶斯公式转换成后验概率
- 根据后验概率大小进行决策分类

通过训练数据集学习联合概率$P(X,Y)$，具体可以学$P(Y=c_k)$与$P(X=x|Y=c_k)$，上面这俩就是要学习的参数

先来认识几个概率

$$\underbrace{P(Y=y|X=x)}_{\text{后验概率}} = \frac{\overbrace{P(X=x|Y=y)}^{\text{似然概率}}\overbrace{P(Y=y)}^{\text{先验概率}}}{\underbrace{P(X=x)}_{\text{证据概率}}}$$

条件独立性假设：假设特征之间是相互独立的

$$\begin{aligned} P(X=x|Y=c_k) &= P(X^{(1)}=x^{(1)},\cdots,X^{(n)}=x^{(n)}|Y=c_k) \\ &= \prod_{j=1}^{n} P(X^{(j)}=x^{(j)}|Y=c_k) \end{aligned}$$

具体的推导

$$P(Y=c_k|X=x)=\frac{P(X=x|Y=c_k)P(Y=c_k)}{\sum_k P(X=x|Y=c_k)P(Y=c_k)}$$
$$=\frac{P(Y=c_k)\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}|Y=c_k)}{\sum_k P(X=x|Y=c_k)P(Y=c_k)}$$

分母是常数，可忽略，最后的目标为最大化y，即

$$y=f(x)=\argmax_{c_k}P(Y=c_k)\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}|Y=c_k)$$

为何最大化后验概率？下文后验概率最大化的含义给出了证明

用人话说就是根据训练集统计出$P(Y=c_k)$和$P(X=x|Y=c_k)$，然后根据测试样本的特征找到对应的$P(X=x|Y=c_k)$，最后算出不同类别的$P(Y=c_k)$和$P(X=x|Y=c_k)$的乘积，选择结果最大的那个作为测试样本的类别

参数估计

极大似然估计

使用样本频率估计概率

$$P(Y=c_k)=\frac{\sum_{i=1}^{N}I(y_i=c_k)}{N}$$

$$P(X^{(j)}=a_{jl}|Y=c_k)=\frac{\sum_{i=1}^{N}I(X_i^{(j)}=a_{jl},y_j=c_k)}{\sum_{i=1}^{N}I(y_i=c_k)}$$

其中：I是指示函数

贝叶斯估计

可以防止某些特征的取值个数为0导致似然概率为0，加上一个平滑项使最后的乘积结果不为0

$$P_\lambda(Y=c_k)=\frac{\sum_{i=1}^{N}I(y_i=c_k)+\lambda}{N+K\lambda}$$

$$P_\lambda(X^{(j)}=a_{jl}|Y=c_k)=\frac{\sum_{i=1}^{N}I(X_i^{(j)}=a_{jl},y_j=c_k)+\lambda}{\sum_{i=1}^{N}I(y_i=c_k)+S_j\lambda}$$

其中：S_j为特征属性取值总数，K为类别的数量

推导朴素贝叶斯中的概率估计公式

贝叶斯估计

对于先验概率

假设进行 N 次实验，先验概率

$$P(Y=c_k)=\frac{1}{K}$$
$$PK-1=0$$

由频率是概率的极大似然估计

$$P(Y=c_k)=\frac{\sum_{i=1}^{N}I(y_i=c_k)}{N}$$

可得

$$P(Y=c_k)N-\sum_{i=1}^{N}I(y_i=c_k)=0$$

即

$$\lambda(P(Y=c_k)K-1)+P(Y=c_k)N-\sum_{i=1}^{N}I(y_i=c_k)=0$$

可得

$$P(Y=c_k)=\frac{\sum_{i=1}^{N}I(y_i=c_k)+\lambda}{K\lambda+N}$$

对于似然概率

S_j 是第 j 个特征可能取值的个数，这里假设每个特征的值是等概率分布，因此概率为 $\frac{1}{S_j}$。

$$P(X^{(j)}=a_{jl}|Y=c_k)=\frac{1}{S_j}$$
$$P(X^{(j)}=a_{jl}|Y=c_k)=\frac{\sum_{i=1}^{N}I(X^{(j)}=a_{jl},Y=c_k)}{\sum_{i=1}^{N}I(Y=c_k)}$$

同上

$$\lambda(PS_j-1)+P\sum_{i=1}^{N}I(Y=c_k)-\sum_{i=1}^{N}I(X^{(j)}=a_{jl},Y=c_k)=0$$

可得

$$P(X^{(j)}=a_{jl}|Y=c_k)=\frac{\lambda+\sum_{i=1}^{N}I(X^{(j)}=a_{jl},Y=c_k)}{\sum_{i=1}^{N}I(Y=c_k)+S_j\cdot\lambda}$$

后验概率最大化的含义

这里来证明一下朴素贝叶斯是如何从损失函数最小化推出后验概率最大化的。

首先我们写出期望风险的公式：

$$\begin{aligned} R_{exp} &= \int\int L(y, f(\vec{x}))P(\vec{x}, y)d\vec{x}dy \\ &= \int_x \int_y L(y, f(\vec{x}))P(y|\vec{x})dyP(\vec{x})d\vec{x} \\ &= \mathbb{E}_x[\int_y L(y, f(\vec{x}))P(y|\vec{x})dy] \\ &= \mathbb{E}_x[\sum_{k=1}^{K} L(c_k, f(\vec{x}))P(c_k|\vec{x})] \end{aligned}$$

也就是对于每个 x，我们对 $L(y, f(\vec{x}))P(y|\vec{x})$ 进行最小化：

$$\begin{aligned} f(x) &= \arg\min_{y\in Y} \sum_{k=1}^{K} L(c_k, y)P(c_k|X=x) \\ &= \arg\min_{y\in Y} \sum_{k=1}^{K} P(y \neq c_k|X=x) \\ &= \arg\min_{y\in Y}(1 - P(y=c_k|X=x)) \quad \text{所有概率和为1，等价为1减去预测正确的概率} \\ &= \arg\max_{y\in Y} P(y=c_k|X=x) \end{aligned}$$

决策树

理想的决策树：

- 叶结点数最少
- 叶结点深度最小
- 叶结点数最小且叶结点深度最小

防止决策树过拟合：剪枝、强制决策树最大深度、规定叶结点最少样本数

熵：表示随机变量的不确定性度量

$$I(a_i) = p(a_i)\log_2 \frac{1}{p(a_i)}$$

设X是一个取有限个值的离散随机变量

$$p(X = x_i) = p_i$$

则随机变量X的熵定义为

$$H(X) = -\sum_i p_i \log_2 p_i$$

信息增益

ID3算法采用的属性选择方式

信息增益其实就是互信息，表示得知特征A的信息而使得不确定性减少的程度，$g(D, A)$定义为集合D的经验熵$H(D)$与特征A给定条件下D的经验条件熵$H(D|A)$之差

$$g(D,A)=H(D)-H(D|A)$$

下面这个是大体流程，刚开始看肯定看不懂，做了题之后就明白了

假设拥有训练数据集D，$|D|$表示其样本容量（样本个数）

设有K个类C_k，$k=1,2,\ldots,$ $|C_k|$为属于类C_k的样本个数

特征A有n个不同的取值$a_1,a_2,\ldots,a_n$，根据特征A的取值，将D划分为n个子集$D_1,D_2,\ldots,D_n$

$|D_i|$为D_i的样本个数，记子集D_i中属于类C_k的样本集合为D_{ik}，$|D_{ik}|$为D_{ik}的样本个数

1. 数据集D的经验熵$H(D)$

$$H(D)=-\sum_{k=1}^{K}\frac{|C_k|}{|D|}\log_2\frac{|C_k|}{|D|}$$

2. 特征A对数据集D的经验条件熵$H(D|A)$

$$H(D|A)=\sum_{i=1}^{N}\frac{|D_i|}{|D|}H(D_i)=\sum_{i=1}^{N}\frac{|D_i|}{|D|}\sum_{k=1}^{K}\frac{|D_{ik}|}{|D_i|}\log_2\frac{|D_{ik}|}{|D_i|}$$

3. 信息增益

$$g(D,A)=H(D)-H(D|A)$$

信息增益最大的特征为最优特征！！！！

信息增益比

C4.5算法采用的属性选择方式

特征A对于训练数据集D的信息增益比定义为信息增益与训练数据集关于特征值A的熵之比

$$g_k(D,A)=\frac{g(D,A)}{H_A(D)}$$

其中：- n是特征值A的取值个数

$$H_A(D)=-\sum_{i=1}^{n}\frac{|D_i|}{|D|}\log_2\frac{|D_i|}{|D|}$$

信息增益比最大的特征为最优特征！！！！

基尼指数

CART算法采用的属性选择方式

要注意的是：CART算法生成的是二叉树，若一个特征有多种属性，你只能划分是属性A和不是属性A两种情况

- 目标变量离散：生成分类树
- 目标变量连续：生成回归树

使用基尼指数选择最优特征，并决定该特征的**最优二值切分点**

对于给定的样本集合D

$$Gini(D) = 1 - \sum_k (\frac{|C_k|}{|D|})^2$$

若样本集合根据特征A被划分为D_1、D_2两个部分，那么在特征A条件下，集合D的基尼指数定义为：

$$Gini(D, A) = \frac{|D_1|}{|D|}Gini(D_1) + \frac{|D_2|}{|D|}Gini(D_2)$$

基尼指数表示不确定性，基尼指数越大，集合不确定性越大，因此我们要选择基尼指数小的特征进行划分

Logistic回归

Logistic回归是广义线性模型，决策边界是线性的

二项Logistic回归

$$P(Y = 1|X) = \frac{e^{w \cdot x + b}}{1 + e^{w \cdot x + b}}$$
$$P(Y = 0|X) = \frac{1}{1 + e^{w \cdot x + b}}$$

事件发生比：发生与不发生概率之比

$$\frac{P}{1 - P}$$

对数几率

$$\log \frac{P}{1 - P}$$

模型参数估计

使用极大似然估计实现

对于n个观测事件$\{(x_i, y_i)\}_{i=1}^N, x_i \in R^n, y_i \in \{0, 1\}$

已知

$$P(Y = 1|x) = \pi(x) \quad P(Y = 0|x) = 1 - \pi(x)$$

可得到似然函数

$$\begin{aligned} L(\pi(x)|x, y) &= P(X = x, Y = y|\pi(x)) \\ &= \prod_{i=1}^N P(X = x_i)P(Y = y_i|X = x_i, \pi(x)) \\ &\propto \prod_{i=1}^N [\pi(x_i)^{y_i}][1 - \pi(x_i)]^{1-y_i} \end{aligned}$$

进行对数化

$$\begin{aligned}\ln L(w) &= \sum_{i=1}^{N} y_i \ln \pi(x_i) + (1-y_i)\ln \pi(x_i) \\ &= \sum_{i=1}^{N} y_i \ln \frac{\pi(x_i)}{1-\pi(x_i)} + \ln\left(1-\pi(x_i)\right) \\ &= \sum_{i=1}^{N} y_i (w \cdot x_i) - \ln\left(1+e^{w\cdot x_i}\right)\end{aligned}$$

对此函数使用梯度下降求极大值，得到w估计值

$$\nabla_w(-\ln L(w)) = \sum_{i=1}^{N} \frac{x_i e^{w\cdot x_i}}{1+e^{w\cdot x_i}} - y_i x_i$$

当$y \in \{-1, 1\}$时

似然函数为

$$L(w) = -\prod_{i=1}^{N} \frac{1}{1+e^{-y_i w x_i}}$$

负对数似然为

$$-\ln L(w) = \sum_{i=1}^{N} \ln\left(1+e^{-y_i w x_i}\right)$$

梯度为

$$\nabla_w(-\ln L(w)) = \sum_{i=1}^{N} \frac{-y_i x_i e^{-y_i w x_i}}{1+e^{-y_i w x_i}}$$

支持向量机

定义在**特征空间上的间隔最大线性分类器**，还包括核函数，使它成为实质上的非线性分类器

不同分类： - 线性可分支持向量机：硬间隔最大化。找到一个超平面，完全正确地将所有样本点分开 - 线性支持向量机：软间隔最大化。找到一个超平面，尽可能正确地将数据点分开，且间隔最大 - 非线性支持向量机：核技巧+软间隔最大化

线性可分支持向量机

给定线性可分训练数据集，通过间隔最大化求得超平面以及分类决策函数

$$f(x) = sign(w^* \cdot x + b^*)$$

函数间隔（成比例改变w和b，超平面没改变，函数间隔却改变）

超平面(w, b)关于样本点(x_i, y_i)的函数间隔为

$$\hat{\gamma}_i = y_i(w \cdot x_i + b)$$

超平面(w,b)关于训练数据集的函数间隔为$\hat{\gamma}_i$的最小值

$$\hat{\gamma} = \min_{i=1,\cdots,N} \hat{\gamma}_i$$

几何间隔（点到直线的距离）

超平面(w,b)关于样本点(x_i,y_i)的几何间隔为

$$\gamma_i = y_i(\frac{w}{||w||} \cdot x_i + \frac{b}{||w||})$$

超平面(w,b)关于训练数据集的几何间隔为γ_i的最小值

$$\gamma = \min_{i=1,\cdots,N} \gamma_i$$

可得到对应关系

$$\gamma_i = \frac{\hat{\gamma}_i}{||w||} \quad \gamma = \frac{\hat{\gamma}}{||w||}$$

核心：正确划分训练数据集并且使得分离超平面的几何间隔最大

学习的最优化问题：

$$\max_{w,b} \gamma \, \mathrm{s.t.} \quad y_i(\frac{w}{||w||} \cdot x_i + \frac{b}{||w||}) \geq \gamma$$

可等价为：

$$\max_{w,b} \frac{\hat{\gamma}}{||w||} \, \mathrm{s.t.} \quad y_i(w \cdot x_i + b) \geq \hat{\gamma}$$

取$\hat{\gamma} = 1$，最大化$\frac{1}{||w||}$和最小化$\frac{1}{2}||w||^2$等价

$$\min_{w,b} \frac{1}{2}||w||^2 \, \mathrm{s.t.} \quad y_i(w \cdot x_i + b) - 1 \geq 0$$

上述的最优化问题是一个凸优化问题，使用拉格朗日对偶法进行求解。求解对偶问题。

构造拉格朗日函数

$$L(w,b,\alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{N} \alpha_i y_i (w \cdot x_i + b) + \sum_{i=1}^{N} \alpha_i$$

α为拉格朗日乘子向量，**有几个样本他就有几个维度**

$$\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_N)^T$$

原始问题

$$\min_{w,b} \max_{\alpha} L(w,b,\alpha)$$

对偶问题

$$\max_{\alpha} \min_{w,b} L(w,b,\alpha)$$

对L求梯度

$$\nabla_w L(w,b,\alpha) = w - \sum_{i=1}^{N} \alpha_i y_i x_i = 0$$

$$\nabla_b L(w,b,\alpha) = \sum_{i=1}^{N} \alpha_i y_i = 0$$

带入原函数

$$\begin{aligned} \theta(\alpha) &= L(w^{'}, b^{'}, \alpha^{'}) \\ &= -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) + \sum_{i=1}^{N} \alpha_i \\ \text{s.t.} \quad & \sum_{i=1}^{N} \alpha_i y_i = 0 \end{aligned}$$

问题求解到这里之后就无法进行人工求解了，需交由机器求解，接下来阐述一些概念性的东西

KKT条件

是为解决带有约束的优化问题而提出的一组规则。如果一个点是局部最优解，那么它必须满足这组规则。满足KKT条件的点就是最优解

$$

$$\begin{cases} \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) = 0 & \text{【平稳性条件】} \\ g_i(\mathbf{x}^*) \leq 0, \quad h_j(\mathbf{x}^*) = 0 & \text{【原始可行性条件】} \\ \lambda_i^* \geq 0 & \text{【对偶可行性条件】} \\ \lambda_i^* g_i(\mathbf{x}^*) = 0 & \text{【互补松弛条件】} \end{cases}$$

$$

其中：$g_i(\mathbf{x}^*)$是不等式约束，$h_j(\mathbf{x}^*)$是等式约束

我们需要着重注意**互补松弛条件**

在上面我们求解出来了$\theta(\alpha)$，接着我们要如何求解w和b呢？ > 机器对$\theta(\alpha)$进行求解得到最后的α > > α是一个向量，我们要取里面不为0的α对应的样本取去对w和b进行求解 > > 这是因为当α_i不为0，由互补松弛条件可知，对应的样本x_i满足$y_i(w \cdot x_i + b) - 1 = 0$，当我们有了这个条件，就可以很方便的求解出$w$和$b$ > > 由$y_i(w \cdot x_i + b) - 1 = 0$可以知道，在两边同乘$y_i (y \in -1, 1)$ > >

$$> (w \cdot x_i + b) y_i^2 = y_i > b = y_i - w \cdot x_i >$$

> > 那么w如何求解呢？ > > 由前面的梯度的条件可以知道 > >

$$> w - \sum_{i=1}^{N} \alpha_i y_i x_i = 0 > w = \sum_{i=1}^{N} \alpha_i y_i x_i >$$

> > 也就是说，α_i不为0的样本对最后的w起了贡献，也满足$y_i(w \cdot x_i + b) - 1 = 0$这个条件，从而可以计算出$w$和$b$

到这里就结束了，理解互补松弛条件至关重要，在线性支持向量机中它仍旧起着至关重要的作用

线性支持向量机

在训练数据有一些特异点，不能满足函数间隔大于等于1地约束条件，为了解决这个问题，给每个样本点引入了一个松弛变量$\xi_i \geq 0$，约束条件变为：

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i$$

目标函数变为：

$$\frac{1}{2}||w||^2 + C\sum_{i=1}^{N}\xi_i$$

其中：$C(>0)$为惩罚系数，C地选取会对模型有着不同的影响，具体看后面的复习章节

还是使用拉格朗日乘子法

$$\min_{w,b,\xi}\frac{1}{2}||w||^2 + C\sum_{i=1}^{N}\xi_i \,\mathrm{s.t.} \quad y_i(w \cdot x_i + b) \geq 1 - \xi_i \quad i = 1,2,\cdots,N \xi_i \geq 0 \quad i = 1,2,\cdots,N$$

拉格朗日函数

$$L(w,b,\xi,\alpha,\mu) = \frac{1}{2}||w||^2 + C\sum_{i=1}^{N}\xi_i + \qquad \sum_{i=1}^{N}\alpha_i(1 - \xi_i - y_i(w \cdot x_i + b)) - \sum_{i=1}^{N}\mu_i\xi_i$$

原始问题

$$

{w,b,}{{,}{L(w,b,,,)}}

$$

对偶问题

$$\max_{\alpha,\mu}\min_{w,b,\xi}L(w,b,\xi,\alpha,\mu)$$

求梯度

$$\nabla_w L(w,b,\alpha) = w - \sum_{i=1}^{N}\alpha_i y_i x_i = 0$$

$$\nabla_b L(w,b,\alpha) = \sum_{i=1}^{N}\alpha_i y_i = 0$$

$$\nabla_\xi L(w,b,\alpha) = C - \alpha_i - \mu_i = 0$$

代入原函数

$$\min_{w,b,\xi}L(w,b,\xi,\alpha,\mu) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j(x_i \cdot x_j) + \sum_{i=1}^{N}\alpha_i$$

对$\min_{w,b,\xi} L(w,b,\xi,\alpha,\mu)$求关于$\alpha$的极大

$$ {} -{i=1}^N{_{j=1}^N{_iy_iy_j(x_ix_j)}} + {i=1}^N{_i} \

_{i=1}^N{_iy_i} = 0 \ C - _i - _i = 0 \ _i \ _i $$

最终的解为

$$\alpha^* = [\alpha_1^*, \alpha_2^*, \cdots, \alpha_n^*]$$

当$\alpha_i^* > 0$时，样本为支持向量

若$\alpha_i < C$，那么$\xi_i = 0$（互补松弛条件）

若$\alpha_i = C$时：

$$\begin{cases} 0 < \xi_i < 1 & \text{分类正确} \\ \xi_i = 1 & \text{落在决策边界上} \\ \xi_i > 1 & \text{被误分类} \end{cases}$$

具体为什么，使用互补松弛条件推一下就可以知道了

非线性支持向量机

究其本质，其实就是找到一个映射函数，把$x_i \cdot x_j$，也就是x_i和x_j的内积映射到一个高维空间在高维空间实现线性可分

正定核充要条件

$K(x,z)$为正定核函数的充要条件是$K(x,z)$对应的Gram矩阵

$$K = \begin{bmatrix} K(x_1,x_1) & \cdots & K(x_1,x_m) \\ \vdots & \ddots & \vdots \\ K(x_m,x_1) & \cdots & K(x_m,x_m) \end{bmatrix}$$

半正定（特征值≥ 0）

合页损失函数的推导

在软间隔支持向量机里面，引入松弛变量 ξ_i。

目标函数为：

$$L = \frac{1}{2}||w||^2 + C\sum_{i=1}^{N}\xi_i$$

约束条件：

$$\begin{aligned} y_i(w^T x_i + b) &\geq 1 - \xi_i \\ \xi_i &\geq 0 \end{aligned}$$

- 当 $y_i(w^T x_i + b) \geq 1$ 时，说明样本被正确分类且间隔足够大
  - 此时 $1 - y_i(w^T x_i + b) \leq 0$，而约束条件为 $\xi_i \geq 1 - y_i(w^T x_i + b)$，即 ξ_i 需要大于一个负数且大于 0。
  - 为了让目标函数最小，令 $\xi_i = 0$。
- 当 $y_i(w^T x_i + b) < 1$ 时，说明样本**间隔不足**或者**被误分类**

- 此时 $1-y_i(w^Tx_i+b)>0$，同上面的推导，约束条件为 $\xi_i \geq 1-y_i(w^Tx_i+b)$，即 ξ_i 需要大于一个正数且大于 0。
- 为满足约束且使目标函数最小，$\xi_i = 1-y_i(w^Tx_i+b)$。

综上所述：

$$\xi_i = \max\left(0, 1-y_i(w^Tx_i+b)\right)$$

代入目标函数并令 $\lambda = \frac{1}{2C}$ 可得合页损失函数：

$$\sum_{i=1}^{N}[1-y_i(w^Tx_i+b)]_+ + \lambda||w||^2$$

其中

$$[z]_+ = \begin{cases} z, & z \geq 0 \\ 0, & z < 0 \end{cases}$$

提升方法

可分为Boosting和Bagging

- Boosting：串行学习，加权采样，加权投票，降低方差
- Bagging：并行学习，Boostrap采用，平等投票，降低偏差

Adaboost

使用的损失函数是指数损失，且使用了前向分布算法，属于加法模型。也就是说每次迭代，弱分类器会遍历每一个特征，寻找误差最小的特征进行样本的重新加权

算法流程

1. 初始化训练数据初始权重分布

$$D_1 = (w_{11}, \cdots, w_{1i}, \cdots, w_{1n}) \quad w_{1i} = \frac{1}{N}$$

2. 对弱分类器$G_m(x)$计算分类误差

$$e_m = \sum_{i=1}^{N} w_{mi} I(G_m(x_i) \neq y_i)$$

3. 计算弱分类器的权重系数

$$\alpha_m = \frac{1}{2}\ln\frac{1-e_m}{e_m}$$

4. 更新训练数据集权重分布

$$D_{m+1} = (w_{m+1,1}, \cdots, w_{m+1,N})$$

$$w_{m+1,i} = \frac{w_{mi}}{z_m} e^{-\alpha_m y_i G_m(x_i)}$$

其中z_m是规范化因子

$$z_m = \sum_{i=1}^{N} w_{mi} e^{-\alpha_m y_i G_m(x_i)}$$

5. 构造基本分类器的线性组合

$$f(x) = \sum_{m=1}^{M} \alpha_m G_m(x) G(x) = sign(f(x))$$

总的来说就是：给每个样本赋予一个初始权重，用m个弱分类器不断更新每个样本的权重，上一轮被分错的样本当前轮的权重会变大

其它的不是考察重点，感兴趣看看课件吧

EM算法

EM算法的推导

它的推导和 VAE（变分自编码器）以及 Diffusion Model（扩散模型）很像，都是由于原函数难以求解，去找到它的下界函数，对下界函数求得最优解，达到优化原函数的目的。感兴趣的可以看一下 VAE 和 Diffusion Model 的推导。

首先要了解 EM 算法的核心：**EM 算法是计算每个潜在变量 Z 的后验概率，然后通过加权平均，也就是求期望的方式得到一个近似的估计，最后我们要找到一个能让这个近似估计最大的参数 θ**。

先来认识一下 **KL 散度**：

$$KL(q||p) = \sum_{Z} q(Z) \log \frac{q(Z)}{p(Z)}$$

描述的是 q 分布和 p 分布之间的距离，因此取值是大于等于 0 的。

首先要明确，EM 算法的核心是在每一步 **最大化 Q 函数**：

$$\begin{aligned} Q(\theta, \theta^{(i)}) &= \mathbb{E}_Z(\log P(Y, Z|\theta)|Y, \theta^{(i)}) \\ &= \sum_{Z} \log P(Y, Z|\theta) P(Z|Y, \theta^{(i)}) \end{aligned}$$

现在来看如何推导出上面这个 Q 函数。

对于 EM 算法，我们希望最大化观测数据 Y 关于参数 θ 的对数似然函数：

$$\begin{aligned} L(\theta) = \log P(Y|\theta) &= \log \sum_{Z} P(Y, Z|\theta) \\ &= \log \sum_{Z} q(Z) \frac{P(Y, Z|\theta)}{q(Z)} \end{aligned}$$

使用 **Jensen 不等式**：

$$\begin{aligned}\log \sum_Z q(Z) \frac{P(Y,Z|\theta)}{q(Z)} &\geq \sum_Z q(Z) \log \frac{P(Y,Z|\theta)}{q(Z)} \\ &= \sum_Z q(Z) \log P(Y,Z|\theta) - \sum_Z q(Z) \log q(Z)\end{aligned}$$

记上式为：

$$\mathcal{L}(q,\theta) = \sum_Z q(Z) \log P(Y,Z|\theta) - \sum_Z q(Z) \log q(Z)$$

因为

$$\log P(Y,Z|\theta) = \log P(Z|Y,\theta) P(Y|\theta)$$

因此有

$$\begin{aligned}\mathcal{L}(q,\theta) &= \sum_Z q(Z) \log P(Y,Z|\theta) - \sum_Z q(Z) \log q(Z) \\ &= \sum_Z q(Z) \log P(Z|Y,\theta) P(Y|\theta) - \sum_Z q(Z) \log q(Z) \\ &= \sum_Z q(Z) (\log P(Z|Y,\theta) + \log P(Y|\theta)) - \log q(Z) \\ &= \log P(Y|\theta) \sum_Z q(Z) + \sum_Z q(Z) \log \frac{P(Z|Y,\theta)}{q(Z)} \\ &= \log P(Y|\theta) - KL(q(Z)||P(Z|Y,\theta))\end{aligned}$$

最终等式

$$\log P(Y|\theta) = KL(q(Z)||P(Z|Y,\theta)) + \mathcal{L}(q,\theta)$$

且

$$\log P(Y|\theta) \geq \mathcal{L}(q,\theta)$$

老师上课的时候没有说 $q(Z)$ 为什么要取 $P(Z|Y,\theta^i)$，这里解释一下：

> **在 E 步的时候**，理论上我们可以取任意的 $q(Z)$ 来构造下界 $\mathcal{L}(q,\theta)$，但是为了让当前参数的似然函数与下界相等，我们要让
>
> $$q(Z) = P(Z|Y,\theta^i)$$
>
> 为什么要这样？ 这样选择能让在参数 $\theta^{(i)}$ 处，下界紧贴似然函数，即：
>
> $$\log P(Y,Z|\theta^{(i)}) = \mathcal{L}(q,\theta^{(i)})$$
>
> （此时 KL 散度为 0）
>
> 当下界紧贴似然函数时，提升下界会对似然函数产生最大程度的提升，收敛速度最快。
>
> 如果下界很松（KL 散度很大），即使提升了下界，似然函数可能只有很小的增长。
>
> **注意**：选择 $q(Z) = P(Z|Y,\theta^{(i)})$ 只是为了在 E 步的当前点 $\theta^{(i)}$ 使得 KL 散度为零、下界紧贴似然函数，从而让后续优化更高效，一旦进入 M 步，参数 θ 开始更新，真实后验 $P(Z|Y,\theta)$ 随之改变，KL 散度立即变为正数，下界与似然函数

> 不再紧贴。

因此

$$\begin{aligned}\log P(Y|\theta) &\geq \mathcal{L}(q,\theta)\\ &= \sum_Z P(Z|Y,\theta^{(i)})\log P(Y,Z|\theta) + const\\ &= \underbrace{\mathbb{E}_Z(\log P(Y,Z|\theta)|Y,\theta^{(i)})}_{\text{Q函数}} + \underbrace{\sum_Z P(Z|Y,\theta^{(i)})\log P(Z|Y,\theta^{(i)})}_{\text{常数}}\end{aligned}$$

单调性的证明

在 E 步的时候，取

$$q(Z) = P(Z|Y,\theta)$$

做完 M 步之后，新下界大于旧下界

$$\mathcal{L}(q,\theta^{(i+1)}) \geq \mathcal{L}(q,\theta^{(i)})$$

旧似然等于旧下界

$$\log P(Y,Z|\theta^i) = \mathcal{L}(q,\theta^{(i)})$$

新似然做完 M 步后 KL 散度大于 0，因此

$$\log P(Y,Z|\theta^{i+1}) \geq \mathcal{L}(q,\theta^{(i+1)})$$

最后有

$$\log P(Y,Z|\theta^{i+1}) \geq \mathcal{L}(q,\theta^{(i+1)}) \geq \mathcal{L}(q,\theta^{(i)}) = \log P(Y,Z|\theta^i)$$

保证了 **EM 算法的单调性**。

高斯混合模型

高斯分布

$$\phi(y|\theta_k) = \frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{(x-\mu_k)^2}{2\sigma_k^2}\right)$$

用多个高斯分布的加权组合描述复杂数据的分布

$$\begin{aligned}&P(y|\theta) = \sum_{k=1}^K \alpha_k \phi(y|\theta_k)\\ &\text{s.t.}\quad \alpha_k \geq 0\ (k=1,\dots,K),\quad \sum_{k=1}^K \alpha_k = 1\end{aligned}$$

Q 函数的定义

$$Q(\theta,\theta^{(i)}) = \mathbb{E}_\gamma[\log P(y,\gamma|\theta)|y,\theta^{(i)}]$$

观测数据是由高斯混合模型产生的，模型参数为

$$\theta = (\alpha_k, \mu_k, \sigma_k^2)$$

隐变量定义为

$$\gamma_{jk} = \begin{cases} 1 & \text{第 } j \text{ 个观测样本来自第 } k \text{ 个高斯模型} \\ 0 \end{cases}$$

现在来推导完全数据的似然函数，因为后续会使用这个代入 Q 函数

$$\begin{aligned} P(y, \gamma|\theta) &= \prod_{j=1}^{N} p(y_j, \gamma_{j1}, \gamma_{j2}, \ldots, \gamma_{jk}|\theta) \\ &= \prod_{j=1}^{N} \prod_{k=1}^{K} [\alpha_k \phi(y_j|\theta_k)]^{\gamma_{jk}} \end{aligned}$$

令

$$n_k = \sum_{j=1}^{N} \gamma_{jk}$$

有

$$\prod_{j=1}^{N} \prod_{k=1}^{K} [\alpha_k \phi(y_i|\theta_k)]^{\gamma_{jk}} = \prod_{k=1}^{K} \alpha_k^{n_k} \prod_{j=1}^{N} [\phi(y_j|\theta_k)]^{\gamma_{jk}}$$

代入到 Q 函数中

$$\begin{aligned} Q(\theta, \theta^{(i)}) &= \mathbb{E}_\gamma \{ \sum_{k=1}^{N} \{ n_k \log \alpha_k + \sum_{j=1}^{N} \gamma_{jk} [\log \phi(y_j|\theta_k)] \} | y, \theta^{(i)} \} \\ &= \mathbb{E}_\gamma \{ \sum_{k=1}^{N} \{ n_k \log \alpha_k + \sum_{j=1}^{N} \gamma_{jk} [(\log \frac{1}{\sqrt{2\pi}} - \log \sigma_k - \frac{(y_j - \mu_k)^2}{2\sigma_k^2})] \} | y, \theta^{(i)} \} \end{aligned}$$

我们可以很轻松地发现，真正需要计算的只有 $\mathbb{E}(\gamma_{jk}|y, \theta^{(i)})$，因为其它的都可以通过期望的线性性质直接拆开得到结果

$$\begin{aligned} \mathbb{E}(\gamma_{jk}|y, \theta^{(i)}) &= P(\gamma_{jk} = 1|y, \theta^{(i)}) \\ &= \frac{P(\gamma_{jk} = 1|\theta^{(i)}) P(y_j|\gamma_{jk} = 1, \theta^{(i)})}{P(y|\theta^{(i)})} \\ &= \frac{\alpha_k \phi(y_j|\theta_k)}{\sum_{k=1}^{K} \alpha_k \phi(y_j|\theta_k)} \\ &= \hat{\gamma}_{jk} \end{aligned}$$

并且我们有

$$n_k = \sum_{j=1}^{N} \gamma_{jk}$$

$$\mathbb{E}[n_k|y,\theta^{(i)}] = \sum_{j=1}^{N} \mathbb{E}[\gamma_{jk}|y,\theta^{(i)}] = \sum_{j=1}^{N} \hat{\gamma}_{jk}$$

最后得到 **Q 函数的最终形式**

$$\begin{aligned} Q(\theta,\theta^{(i)}) &= \sum_{k=1}^{N}\{\sum_{j=1}^{N}(\mathbb{E}\gamma_{jk})\log\alpha_k + \sum_{j=1}^{N}(\mathbb{E}\gamma_{jk})[(\log\frac{1}{\sqrt{2\pi}} - \log\sigma_k - \frac{(y_j-\mu_k)^2}{2\sigma_k^2})]|y,\theta^{(i)}\} \\ &= \sum_{k=1}^{N}\{\sum_{j=1}^{N}\hat{\gamma}_{jk}\log\alpha_k + \sum_{j=1}^{N}\hat{\gamma}_{jk}[(\log\frac{1}{\sqrt{2\pi}} - \log\sigma_k - \frac{(y_j-\mu_k)^2}{2\sigma_k^2})]|y,\theta^{(i)}\} \end{aligned}$$

> 解释一下上方公式难以理解的点
>
> - $\mathbb{E}(\gamma_{jk}|y,\theta^{(i)})$ 的值其实就是 $P(\gamma_{jk}=1|y,\theta^{(i)})$，因为前面定义过隐变量的值不是 1 就是 0，这里可以很容易看出
> - $\hat{\gamma}_{jk}$ 其实就是样本点 y_j 对高斯分布 k 的隶属度
> - $P(y|\theta^{(i)})$ 不是条件概率，而是在参数为 $\theta^{(i)}$ 的情况下的似然函数，而 y 可以来自不同的高斯分布，因此写成求和的形式
> - $P(\gamma_{jk}=1|\theta^{(i)})$ 就是在参数为 $\theta^{(i)}$ 的情况下数据点来自第 k 个分量的概率，也就是权重 α_k

最后对模型参数 α,μ,σ 进行更新

$$\hat{\alpha}_k = \frac{n_k}{N} = \frac{\sum_{j=1}^{N}\hat{\gamma}_{jk}}{N}$$

$$\hat{\mu}_k = \frac{\sum_{j=1}^{N}\hat{\gamma}_{jk}y_j}{\sum_{j=1}^{N}\hat{\gamma}_{jk}}$$

$$\hat{\sigma}_k = \frac{\sum_{j=1}^{N}\hat{\gamma}_{jk}(y_j-\mu_k)^2}{\sum_{j=1}^{N}\hat{\gamma}_{jk}}$$

聚类方法

核心是相似度或者距离，因为是通过这俩个指标进行聚类的

闵可夫斯基距离

$$d_{ij} = (\sum_{k=1}^{m}|x_{ki}-x_{kj}|^p)^{\frac{1}{p}}$$

距离越大，相似度越小

- $p=1$：曼哈顿距离
- $p=2$：欧氏距离
- $p=\infty$：切比雪夫距离

相关系数

$$r_{ij}=\frac{\sum_{k=1}^{m}(x_{ki}-\bar{x_i})(x_{kj}-\bar{x_j})}{[\sum_{k=1}^{m}(x_{ki}-\bar{x_i})^2\sum_{k=1}^{m}(x_{kj}-\bar{x_j})^2]^{\frac{1}{2}}}$$

其中：m为样本维度，且

$$\bar{x_i}=\frac{1}{m}\sum_{k=1}^{m}x_{ki}\quad \bar{x_j}=\frac{1}{m}\sum_{k=1}^{m}x_{kj}$$

相关系数的绝对值越接近1，样本越相似；越接近0，越不相似

夹角余弦

$$S_{ij}=\frac{\sum_{k=1}^{m}x_{ki}x_{kj}}{[\sum_{k=1}^{m}x_{ki}^2\sum_{k=1}^{m}x_{kj}^2]^{\frac{1}{2}}}$$

值越接近1，样本越相似；越接近0，越不相似

硬聚类：一个样本只能属于一个类，类交集为空 **软聚类**：一个样本可属于多个类，类交集不为空

类的均值、类中心

$$\bar{x_G}=\frac{1}{n_G}\sum_{i=1}^{n_G}x_i$$

类直径：任意两样本之间的最大距离

类之间的距离：

- 最短距离（单连接）：两类样本之间最短距离
- 最长距离（完全连接）：两类样本之间最长距离
- 中心距离：类中心之间的距离
- 平均距离：任意两样本之间距离的平均值

层次聚类

属于硬聚类

- 聚合聚类（自下而上）：样本刚开始独占一类，不断合并
- 分裂聚类（自上而下）：所有样本为一类，不断分裂

聚合聚类过程

1. 计算两两样本之间的欧氏距离，记作矩阵D
2. 构造n个类，每个类只包含一个样本
3. 合并类间距（可以选取不同的类间距离）最小的两类，构建一个新的类

按照上述步骤不断迭代，生成层次聚类树，最后想要聚成几个类可以在这个聚类树里面找

k-means聚类

属于硬聚类，容易陷入局部最优

样本之间的距离采用**欧氏距离的平方**

损失函数为样本与其所属类中心距离的总和

$$C^* = \argmin_C w(C) = \argmin_C \sum_{l=1}^{k} \sum_{C(i)=l} ||x_i - \bar{x}_l||^2$$

步骤

1. 初始化，$t = 0$，随机选取k个样本点作为初始聚类中心

$$m^{(0)} = (m_1^{(0)}, \cdots, m_k^{(0)})$$

2. 对样本聚类，计算每个样本到类中心的聚类，指派到最近的类中
3. 计算新的类中心，计算各个类样本均值作为新的类中心
4. 重复上述步骤直到满足停止条件

时间复杂度为$O(nmk)$，n是样本个数，m是样本维度，k是类别个数

k初值选择

可以使用kmeans++算法或者肘部法则寻找最优k值

SVD

奇异值分解

$$A_{m\times n} = U_{m\times m}\Sigma_{m\times n}V_{n\times n}^T$$

其中：

$$\Sigma = \text{diag}(\sigma_1, \sigma_2, \cdots, \sigma_n)\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_\rho \geq 0 \quad \rho = \min(m, n)$$

- σ_i为A的奇异值，是AA^T和A^TA特征值(≥ 0)的平方根
- A的列向量为左奇异向量，是AA^T特征向量
- V的列向量为右奇异向量，是A^TA特征向量
- A不一定是方阵

奇异值分解的性质：

1. 奇异值唯一，U和V不唯一，奇异值分解不唯一
2. A和Σ的秩相等，等于正奇异值σ_i的个数
3. $A_{m\times n}$满足如下公式

$$Av_j = \sigma_j u_j \quad A^Tu_j = \sigma_j v_j \quad A^Tu_j = 0 \quad Av_j = 0$$
$$1 \leq j \leq r \quad 1 \leq j \leq r \quad j \geq r+1 \quad j \geq r+1$$

4. 前r个右奇异向量构成$R(A^T)$（A^T的列空间）的一组标准正交基，后$n-r$个右奇异向量构成$N(A)$（零空间）的一组标准正交基。前r个作奇异向量构成$R(A)$（A的列空间）的一组标准正交基，后$n-r$个右奇异向量构成$N(A^T)$标准正交基。

紧奇异值分解

$$A = U_{m\times r}\Sigma_{r\times r}V_{n\times r}^T \quad rank(A) = r < \min(m, n)$$

其中：$U_{m\times r}$为U的前r列，$\Sigma_{r\times r}$为Σ的前r个对角线元素，$V_{n\times r}$为V的前r列

且

$$rank(\Sigma_{r\times r}) = rank(A)$$

截断奇异值分解

Σ比原矩阵低秩

$$

A U_{m k}{k k}V{n k}^T rank(A) = r < k < r

$$

其中：$U_{m\times k}$为U的前k列，$\Sigma_{k\times k}$为Σ的前k个对角线元素（**只取最大的前k个奇异值**），$V_{n\times k}$为V的前k列

矩阵的最优近似

矩阵的外积展开式

$u_i v_i^T$矩阵的外积

$$A = U\Sigma V^T = \sigma_1 u_1 v_1^T + \cdots + \sigma_n u_n v_n^T$$

可简写成

$$A = \sum_{k=1}^{n} \sigma_k u_k v_k^T$$

其中n是矩阵的秩

而

$$A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T = U_k \Sigma_k V_k^T$$

F范数

$$||A||_F = (\sum_i \sum_j a_{ij}^2)^{\frac{1}{2}} = \mathrm{tr}(A^T A)^{\frac{1}{2}}$$

使用F范数计算低秩近似误差

$$||A - A_k||_F^2 = \sigma_{k+1}^2 + \sigma_{k+2}^2 + \cdots + \sigma_r^2$$

SVD的几何解释

从右到左看$A = U\Sigma V^T$，先V^T，再Σ，最后U

- V的列向量构成R^n空间的一组标准正交基，表示R^n空间的正交坐标系的旋转或反射变换
- Σ的对角线元素是一组非负实数，表示R^n空间中的原始正交坐标系坐标轴的$\sigma_1, \sigma_2, \cdots, \sigma_n$倍缩放变换
- U的列向量构成R^m空间的一组标准正交基，表示R^m空间的正交坐标系的旋转或反射变换

PCA

是一种降维方法，利用正交变换把线性相关变量转换成少数几个由线性无关变量来表示数据

首先定义$\vec{x}=(x_1,x_2,\cdots,x_m)^T$是m维随机变量，均值向量是$\mu$

$$\mu=E(\vec{x})=(\mu_1,\mu_2,\cdots,\mu_m)^T$$

通过看上课的PPT可以知道协方差矩阵的定义

$$\Sigma=\mathrm{Cov}(\vec{x},\vec{x})=E[(\vec{x}-\vec{\mu})(\vec{x}-\vec{\mu})^T]$$

然后定义一个线性变换，这个线性变换其实就是α_i去乘以每个样本，得到一个新的坐标的一个维度

> 这里为什么说是一个维度呢，因为你是根据这个线性变换求出来的其实是一个值，这个值就代表了变换后坐标的一个维度。也就是说你可以控制y的个数，如果你要降维，y的个数就不要超过原始样本维度。

$$y_i=\alpha_i^T\vec{x}=\alpha_{1i}x_1+\alpha_{2i}x_2+\cdots+\alpha_{mi}x_m\alpha_i^T=(\alpha_{1i},\alpha_{2i},\cdots,\alpha_{mi})$$

对于$y_i=\alpha_i^T\vec{x}$，若满足：

- α_i^T是单位向量，即$\alpha_i^T\alpha_i=1$
- y_i与y_j不相关，即$\mathrm{Cov}(y_i,y_j)=0$
- y_i是与前面求出来的与y_1到y_{i-1}不相关的线性变换中方差最大的，也就是y_i的方差是小于$y_1,y_2,\cdots,y_{i-1}$的

这个时候就可以称y_1到y_m是x的第一主成分到第m主成分

由此可以得到求主成分的方法

即已知$\vec{x}=(x_1,x_2,\cdots,x_m)^T$是m维随机变量，$\Sigma$是$x$的协方差矩阵，其特征值为$\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_m\geq0$，对应的**单位特征向量**分别是$\alpha_1,\alpha_2,\cdots,\alpha_m$

x的第k个主成分是

$$y_i=\alpha_k^T\vec{x}=\alpha_{1k}x_1+\alpha_{2k}x_2+\cdots+\alpha_{mk}x_m$$

方差为

$$\mathrm{var}(y_k)=\alpha_k^T\Sigma\alpha_k=\lambda_k$$

也就是协方差矩阵第k个特征值

其性质有：

1. $\mathrm{Cov}(\vec{y})=\Lambda=\mathrm{diag}(\lambda_1,\cdots,\lambda_m)$
2. $\vec{y}$的方差之和等于随机变量x的方差之和

$$\sum_{i=1}^{m}\lambda_i=\sum_{i=1}^{m}\sigma_{ii}$$

因子负荷量

第k个主成分y_k与变量x_i的相关系数$\rho(y_k,x_i)$，表示 y_k与x_i的相关关系，要注意的是这里的x_i表示的是样本的第i个维度，而不是第i个样本

$$\rho(y_k, x_i) = \frac{\sqrt{\lambda_k}\alpha_{ik}}{\sqrt{\sigma_{ii}}}$$

其中：λ_k是y_k对应的特征值，α_{ik}是主成分y_k对应的特征向量的第i个维度，σ_{ii}是x_i的方差

固定第k个主成分y_k，有如下性质：

$$\sum_{i=1}^{m} \sigma_{ii}\rho(y_k, x_i)^2 = \sum_{i=1}^{m} \lambda_k \alpha_{ik}^2 = \lambda_k \alpha_k^T \alpha_k = \lambda_k$$

固定第i个维度x_i，有如下性质：

$$\sum_{k=1}^{m} \rho(y_k, x_i)^2 = 1$$

在后面对规范化后的随机变量（方差变为1）进行上述计算时，$\sigma_{ii} = 1$

方差贡献率

用于选取主成分个数

$$\eta_k = \frac{\lambda_k}{\sum_{i=1}^{m} \lambda_i}$$

PCA步骤

首先要确定样本是按行分布还是按列分布的，我这里默认样本按列分布，且使用特征值分解进行PCA求解（课本有使用奇异值分解求解PCA的步骤，由于不是考察重点，这里不再赘述）

1. 规范化随机变量：具体说就是算出样本每一个维度的均值和样本方差，对其进行规范化

$$\bar{x_i} = \frac{1}{n}\sum_{j=1}^{n} x_{ij}$$

$$s_{ii} = \frac{1}{n-1}\sum_{j=1}^{n} (x_{ij} - \bar{x_i})^2$$

$$x_{ij} = \frac{x_{ij} - \bar{x_i}}{\sqrt{s_{ii}}}$$

$$A \to B \quad \text{进行规范化}$$

2. 规范化后求协方差矩阵（因为样本已经归一化，也可以叫做相关矩阵）

$$R = \frac{1}{n-1}BB^T \quad \text{样本按列分布} R = \frac{1}{n-1}B^TB \quad \text{样本按行分布}$$

3. 求解相关矩阵的特征向量，求出主成分，并求出他们的因子负荷量，方差贡献率等，按照题目要求进行主成分个数的选取

PCA的一些证明

帮助加深理解

如何得到用x的协方差矩阵求主成分的结论

> 可以转换为下面带约束的最优化问题
>
> $$\max_{\alpha_1} \alpha_1^T \Sigma \alpha_1 \, \text{s.t.} \quad \alpha_1^T \alpha_1 = 1$$
>
> 构建拉格朗日函数
>
> $$L(\alpha_1) = \alpha_1^T \Sigma \alpha_1 - \lambda_1(\alpha_1^T \alpha_1 - 1)$$
>
> 求导
>
> $$\nabla_{\alpha_1} L = 2\Sigma\alpha_1 - 2\lambda_1\alpha_1 = 0$$
>
> 可得到
>
> $$\Sigma\alpha_1 = \lambda_1\alpha_1$$
>
> α_1是Σ的特征向量，又因为$\text{var}(y_k) = \alpha_k^T \Sigma \alpha_k = \lambda_k$
>
> 要使$\text{var}(y_1)$最大，所以λ_1为Σ的最大特征值

另一个例子 > >

$$> \max_{\alpha_2} \alpha_2^T \Sigma \alpha_2 > \text{s.t.} \quad \alpha_2^T \alpha_2 = 1 \quad \alpha_1^T \alpha_2 = 0 >$$

> > 求拉格朗日函数 > >

$$> L(\alpha_2) = \alpha_2^T \Sigma \alpha_2 + \lambda_2(1 - \alpha_2^T \alpha_2) + \beta_2 \alpha_1^T \alpha_2 >$$

> > 求梯度 > >

$$> \nabla_{\alpha_2} L = 2\Sigma\alpha_2 - 2\alpha_2\lambda_2 + \beta_2\alpha_1 = 0 >$$

> > 左乘α_1^T > >

$$> 2\alpha_1^T \Sigma \alpha_2 - 2\alpha_1^T \alpha_2 \lambda_2 + \beta_2 \alpha_1^T \alpha_1 = \beta_2 = 0 >$$

> > 因此 > >

$$> \Sigma\alpha_2 = \lambda_2\alpha_2 >$$

主成分方差的证明 > 已知 > >

$$> y_1 = \alpha_i^T \vec{x} \quad E(\vec{x}) = \mu = 0 >$$

> > 计算协方差矩阵 > >

$$> \Sigma = \text{Cov}(\vec{x}, \vec{x}) = E((\vec{x} - \mu)(\vec{x} - \mu)^T) = E(\vec{x}\vec{x}^T) >$$

> > 计算主成分的均值 > >

$$> \text{E}(y_i) = \alpha_i^T E(\vec{x}) = \alpha_i^T \mu >$$

>> 计算主成分方差 >>

$$
\begin{aligned}
&> \operatorname{var}(y_i) = E[(y_i - E(y_i))^2] \\
> \quad &>= E(y_i^2) - [E(y_i)]^2 \quad > \\
&>= E(\alpha_i^T \vec{x} \vec{x}^T \alpha_i) \\
&>= \alpha_i^T \Sigma \alpha_i >
\end{aligned}
$$

>> 主成分的协方差矩阵也可以算出来 >>

$$
\begin{aligned}
&> \operatorname{Cov}(y_i, y_i) = E(y_i y_i^T) \\
> \quad &>= \operatorname{var}(y_i) \quad > \\
&>= \alpha_i^T \Sigma \alpha_i >
\end{aligned}
$$

向量和矩阵求导

对向量求导

$$
\begin{aligned}
\frac{\partial x^T a}{\partial x} &= a \\
\frac{\partial a^T x}{\partial x} &= a \\
\frac{\partial x^T x}{\partial x} &= 2x \\
\frac{\partial x^T B x}{\partial x} &= (B + B^T)x
\end{aligned}
$$

对矩阵求导

$$
\begin{aligned}
\frac{\partial a^T X b}{\partial X} &= a b^T \\
\frac{\partial a^T X^T b}{\partial X} &= b a^T \\
\frac{\partial a^T X^T a}{\partial X} &= a a^T \\
\frac{\partial a^T X a}{\partial X} &= a a^T \\
\frac{\partial a^T X X^T a}{\partial X} &= X(a b^T + b a^T)
\end{aligned}
$$

PPT

统计学习概论

1. 什么是统计学习

统计学习是计算机基于数据构建概率统计模型并运用模型对数据进行预测和分析

2. 什么是过拟合

过拟合指的是模型在已知的数据上预测得很好，但是在未知的数据上预测得很差

1. 增加更多的训练样本总是可以避免过拟合。[**错误**]

计算Precision和Recall，体温超过38度为阳性患者

No.	1	2	3	4	5	6	7	8	9	10	11	12
Ground Truth	P	P	F	F	F	P	P	F	F	F	F	F
Tempreture	40°C	39°C	38.7°C	38.6°C	38.3°C	38.1°C	37.8°C	37.6°C	37.4°C	37.2°C	37°C	36.6°C

image-20251216104253184

	P	N
T	TP：3	FP：1
F	FN：3	TN：5

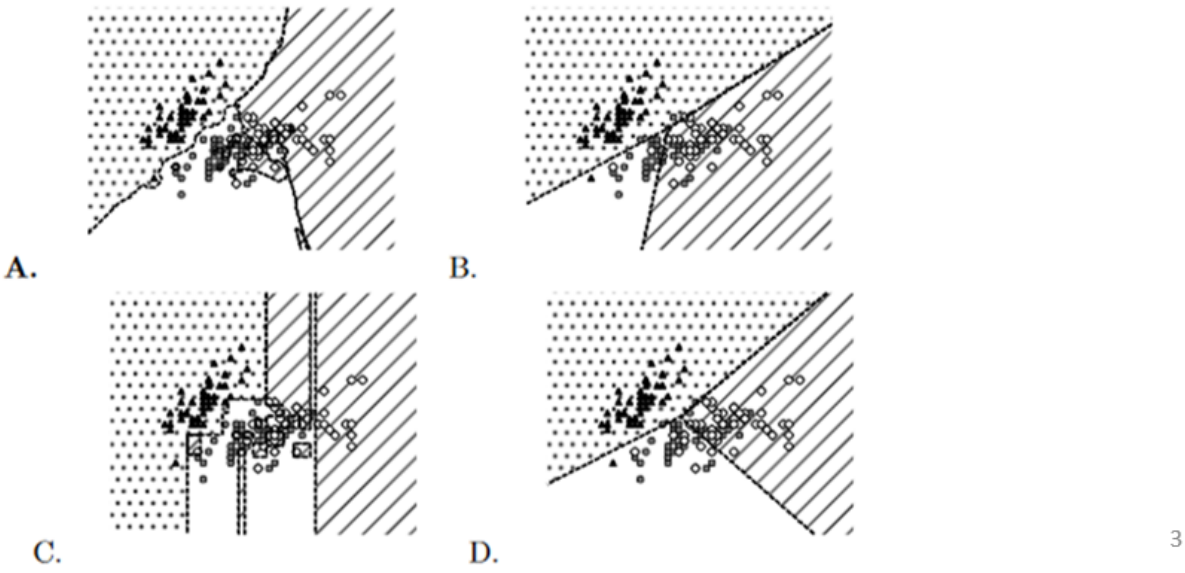
$$Precision = \frac{TP}{TP + FP} = 0.75$$

$$Recall = \frac{TP}{TP + FN} = 0.5$$

knn

1. 1-NN分类器的训练误差为0。**[正确]**
2. 当K=N时，K近邻算法产生的决策边界比1近邻算法更复杂。**[错误]**
3. 选择不同的距离范数不会影响1-NN的决策边界。**[错误]**
4. 最近邻算法是一种参数化方法。**[错误]**

(1 point) Which of the following decision boundaries is most likely to be generated by a k-NN?



3

image-20251216110606101

A，解析：

- 只有A的分类边界是不规则的

朴素贝叶斯

1. 感知机的训练错误率为0。**[错误]**
2. 最近邻分类器的训练错误率为0。**[正确]**
3. 朴素贝叶斯分类器的训练错误率为0。**[错误]**

决策树

1. ID3方法能够保证得到最优的决策树。**[错误]**

下列哪些策略可能减小决策树的过拟合问题：

A.剪枝

B.强制叶节点最小样本数

C.强制决策树的最大深度

D.确保每个叶节点的样本都属于同一类

ABC，解析：

- D：这样子会使得模型为了使叶子节点的样本只有一类，会使得树的深度骤增，使模型过拟合

支持向量机

1. 支持向量机计算$P(y|x)$。**[错误]**
2. 在支持向量机中，非支持向量的α_i的值为0。**[正确]**
3. 在支持向量机中，正例对应的拉格朗日乘子之和等于负例对应的拉格朗日乘子之和。**[正确]**
4. 线性可分SVM的损失函数只考虑误分类的样本。**[错误]**
5. 非线性SVM在训练时不需要设置任何超参数。**[错误]**
6. SVM在处理非线性问题时使用的技术称为：

A. 神经网络 B. 随机森林

C. 核技巧 D. 梯度提升

C

2. 当你想要提高SVM模型的泛化能力，你应该：

A. 增加正则化参数C B. 减少正则化参数C

C. 增加核函数参数γ D. 减少核函数参数γ

BD

提升方法

1. 下列哪个说法是正确的：

A.Adaboost算法可直接用于1NN分类器

B.Adaboost算法运用软间隔线性支持向量机做弱分类器可能得到非线性的分类边界

C.Adaboost算法对测试样本的弱分类器输出运用投票法进行决策

D.Adaboost算法适合任何的一组弱分类器

B，解析：

- A：1NN分类器是强分类器，不适合用在Adaboost里面
- C：是加权投票，不是平均投票
- D：不适合，弱分类器要比随即猜测略好，误差小于0.5

2. 在 Adaboost 中，我们从训练样本上的高斯权重分布开始。**[错误]**

EM算法

EM算法的E步中，我们通常计算什么？ （）

A. 参数的后验概率 B. 缺失数据的期望值

C. 观测数据的期望值 D. 参数的最大似然估计

B，解析：

- 应该是完全数据似然函数在隐变量后验概率分布下的期望

聚类

1. 层次聚类中，距离度量不包括以下哪种 ( )

A. 欧氏距离 B. 曼哈顿距离

C. 余弦相似度 D. 标准差

D

2. 层次聚类算法的类型包括 ( )

A. 凝聚的 B. 分裂的 C. 以上都是 D. 以上都不是

C

3. 层次聚类的结果通常表示为 ( )

A. 一个聚类数 B. 一个聚类划分

C. 一个树状图（Dendrogram） D. 一个概率模型

C

1. 层次聚类方法需要预定义聚类数量。**[错误]**
2. 单链接聚合聚类算法基于两个聚类中点之间的最大距离来合并这两个聚类。**[错误]**
3. K-means算法是一种无监督学习算法 。**[正确]**
4. K-means算法是KNN算法的特例 。**[错误]**
5. K-means算法总是能找到全局最优解，与初始中心选择无关 。**[错误]**
6. K-means算法的K值必须由用户预先指定。**[正确]**

7. K-means算法的时间复杂度是O(nkm)，其中n是数据点的数量，k是簇的数量，m是样本维数。**[正确]**
8. K-means算法可以通过肘部法则来确定K值。**[正确]**

三要素

方法	模型	策略	算法	损失函数
感知机	二分类超平面	极小化误分类点到超平面的距离	随机梯度下降	误分类点到超平面的距离
knn	\	\	\	\
朴素贝叶斯	特征与类别的联合概率	极大似然估计，极大后验概率估计	EM算法	对数似然损失
决策树	分类树、回归树	正则化的极大似然估计	特征选择、生成、剪枝	对数似然损失
逻辑斯特回归	对数线性模型	正则化的极大似然估计	梯度下降，拟牛顿法	逻辑斯蒂损失
支持向量机	分离超平面	最小间隔最大化，极小化带正则的合页损失	SMO算法	合页损失
提升方法	弱分类器的线性组合	极小化加法模型的指数损失	前向分布加法算法	指数损失
EM算法	含隐变量的概率模型	极大似然估计	迭代算法	对数似然损失
层次聚类	聚类树	类内样本距离最小	启发式算法	\
k均值聚类	k中心聚类	样本到类中心的距离之和最小	迭代算法	\
高斯混合模型	高斯混合模型	似然函数最大	EM算法	\
PCA	低维正交空间	方差最大	SVD或者特征值分解	\

我觉得易错的点

1. 感知机对偶是用一个α来存储每个样本的误分类次数乘上学习率，α的维度等于样本个数
2. knn是懒惰学习，是无参数模型，缺点是时间和空间复杂度高，它的模型三要素是距离度量、k值选择、分类决策规则
3. adaboost的训练轮数需要预先指定，训练几轮就有几个弱学习器
4. CART算法是二元划分
5. 用最小训练误差划分决策树会导致过拟合，不能正确反映划分后两个子集各自的数据分布，生成的决策树也易受噪声影响

6. 分裂聚类的时间复杂度是$O(n^2mk)$，层次聚类是层次化的，kmeans是非层次化的
7. 逻辑斯蒂回归是线性分类模型，是广义线性模型，但输入输出不存在线性关系
8. 决策树模型使用的是`if-then`规则，特征是互斥且完备，直接应用是`CLS`算法
9. 奇异值分解的基本定理是奇异值分解对任意实矩阵存在
10. 监督学习的基本策略是经验风险最小化和结构风险最小化
11. SVM的惩罚系数C太大会导致模型过拟合，C太小会导致模型欠拟合
12. SVM是一个求解凸二次规划的问题

考试题型

五选择

六判断

三简答

三到四个计算题

一道综合题（除了第1问都很难）

重点

::: tabs @tab:active 概论 统计学习定义 统计学习对象 统计学习目的 统计学习三要素 每个模型的三要素是什么 比较用三要素 模型分类概率模型和非概率模型(p11-12) 生成模型和判别模型(生成模型关注联合分布 p28有)。损失函数与风险函数 期望损失 损失函数的期望 期望风险 经验风险 结构风险 监督学习定义p6 无监督学习定义p8 过拟合定义 解决办法 混淆矩阵 召回率 精确率 F1值

@tab 监督学习 监督学习 七个方法 感知机收敛定理。模型咋写 损失函数是什么 如何推导 knn决策规则 距离度量k值选择造成的影响 不考kd树 朴素贝叶斯对什么做条件独立性假设 最大似然估计和拉普拉斯平滑 决策树 id3 CART 熵以2为底 决策树剪枝的作用 不考回归树 逻辑斯谛回归要会写最大似然函数 会对其求导 支持向量机原问题 对偶问题 约束 KKT条件 原问题如何变成对偶问题 软间隔svm 合页损失 C变大变小的影响 根据kexi取值范围判断是不是支持向量 核方法 正定核充要矩阵:GRAM矩阵半正定 提示方法概念 弱分类器单独分类正确率要超过0.5 adaboost指数损失

@tab 无监督学习 EM算法要会推导。收敛到局部最大值 鞍点 GMM概念 EM算法两步 层次聚类 单链接还是多链接 不考中心距离 平均距离 例14.1 层次聚类图 kmeans 例14.2 会收敛 不是全局最优 SVD定义 怎么求UsigmaV 例15.5 奇异值分解不唯一 定理15.3 截断奇异值近似误差 F范数 pca性质 yi的特征值和 方差贡献率是什么 相关系数是什么 用拉格朗日乘子法证明要找方差最大的作为变换向量 要会求相关矩阵 规范化变量情况下的性质