No.

Date

学习理论

偏差：Bias ，欠拟合，高偏差

方差：variance ，过拟合，高方差

泛化能力：模型在测试集上的分类效果。

通过 联合界和 Hoeffding不等式，证明，最优化ERM(经验风险最小化)能够来被小训练泛化误差

一致收敛性。

$\varepsilon(\hat{h})$ $\varepsilon(h^*)$

一致收敛定理：训练误差和泛化误差随样本数目的增大而更加接近

最大差距为2γ $\varepsilon(\hat{h}) \le \hat{\varepsilon}(\hat{h})+\gamma \le \hat{\varepsilon}(h^*)+\gamma+\gamma \le \varepsilon(h^*)+\gamma+\gamma$

ERM的定义：$\hat{h} = \arg\min_{h\in H} \hat{\varepsilon}(h)$

训练误差最小的假设h。

$$\hat{\varepsilon}(h) = \frac{1}{m}\sum_{i=1}^{m} I(h_\theta(x^{(i)}) \neq y^{(i)})$$

训练误差。

- VC维 → VC维解释SVM：SVM算法会自动寻找一个具有较小VC维的假设，这样降低了VC维
- ERM的真正意义。
- 模型选择：
  - 交叉检验会
    - 保留交叉检验会
    - K重交叉检验会
  - 特征选择
  - 特征提取

  通过减少模型参数，简化模型来降低过拟合的发生

正则化：在文本分类问题上，特征数目有时会远远大于样本数目，这样就会产生过拟合，添加惩罚项是解决该问题的一个不错方法

No.

Date

在线学习：

对感知器算法，若正负样本对分，那么在线学习算法也是收敛的。

当算法遇到瓶颈时，该选择什么样的方向来对算法进行改进？（算法诊断）

偏/方差分析

一般的，高方差针对过拟合问题，即训练误差很小，但泛化误差很大。

高偏差针对模型不合适问题，如特征数过少，表现是，训练误差和泛化误差都很大。

收敛与目标函数是否一致正确的判断

误差分析

销蚀分析

No.

Date

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

监督学习：

线性回归 —最小二乘法→ $J(\theta) = \frac{1}{2}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$ —求解方法→ 梯度下降 / 正规方程组

正规方程组：$\theta = (X^TX)^{-1}X^Ty$

解决非线性问题

欠拟合问题

过拟合

欠拟合

解决方法 → 非参数学习算法(如：局部加权回归，loess)

本质上可以解释，为什么选择最小二乘作为计算参数的指标，使得预测出的值和y值之间的面积、的平方最小。

分类问题算法：(线性分类)

Logistic回归 → 解决为是离散问题

↳ 可以使用以下方法拟合：梯度上升 / 牛顿方法(更快，但每次都要计算Hessian矩阵)

伯努利分布 ⟹ logistic模型

连续分布 ⟹ 线性模型

指数分布族：若 $P(y|x;\theta)$，y属实数，满足高斯分布，得到基于最小二乘的线性回归

y属{0,1}，满足伯努利分布，得到logistic回归

$$P(y;\eta) = b(y)\exp(\eta^T T(y) - a(\eta))$$

广义线性模型(GLM)

下假设：1、$y|x;\theta \sim ExponentialFamily(\eta)$，即假设对因预测的变量y，在给定x，以θ为参数的条件概率，属于以η作为自然参数的指数分布族

2、给定x，目标是求出以x为条件的T(y)的期望 $E(T(y|x))$，即让学习算法输出：$h(x) = E(T(y|x))$

3、$\eta = \theta^Tx$，即自然参数和输入特征x之间线性相关，关系由θ决定，可以η是实数时才有意义。

若η是一个向量，则 $\eta_i = \theta_i^Tx$

(选定一个指数分布族后，相当于GLM：求 $h_\theta(x)$，相当于求θ)

广义线性回归模型

判别学习算法：对P(y|x)建模

生成学习算法：对P(x|y)建模

→ 高斯判别(GDA) → GDA的假设强 ⟷ Logistic回归的假设弱，更鲁棒.

→ 朴素贝叶斯 → 条件独立假设 → 新样本属性缺失问题，Laplace平滑解决

非线性分类算法

1. 神经网络

2. SVM(支持向量机) → 基础 → 最大间隔分类器.

为了找快速的解决最优间隔问题

→ 应用拉格朗日对偶性质

原始问题和对偶问题 → 最优解等价条件

→ KKT

→ 得到最优间隔分类器的对偶形式

求解时发现，目标函数中含有内积形式.

→ 引入核方法 (得到完完全全的SVM求解问题)

数 最优间隔分类器分：线性：一般不太符合实际

软间隔分类器 ← 应用 ← 非线性

方法，得到最终问题

→ 应用序列最小化算法(SMO)求解

→ END.

简化的版SMO的思想.

求出a的最优解的方法：

坐标上升法

(之前介绍过梯度下降和牛顿法)

No.

Date

无监督学习

K-Means：

混合高斯分布（MoG）

如：MoG中数据属于哪个分组，可以看成z

一般来说，任意形状的概率分布都可以用多个高斯分布函数去近似，因而MoG应用广泛。

利用EM算法求解MoG

1、设置初始θ
2、E-Step，估计隐含变量z
3、M-Step，根据z，重新估计θ
4、重复2、3，直到参数变化小于阈值

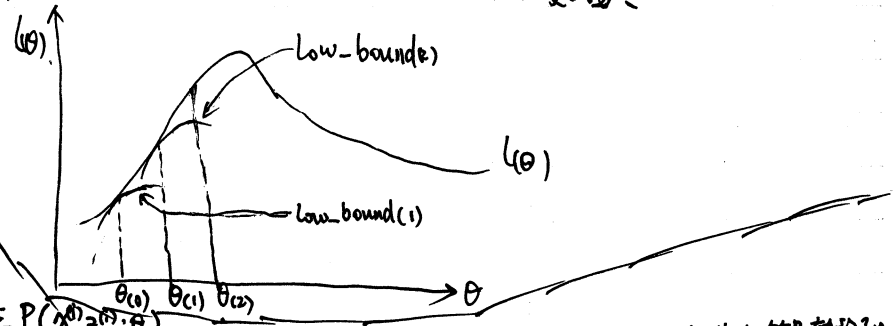
→ 相关前提：Jensen不等式：若f为凸函数，即 $f''(x)\geq 0$，注意，并不一定要对为号

但若存在二阶导数，则必须大于等于0，再令x为随机变量，则有：$f(E[x]) \leq E[f(x)]$

进一步，若二阶导数恒大于0，则不等式等号成立当且仅当 $x=E[x]$，即x是固定值。

若二阶导数≤0方向逆转，则不等式的不等号方向逆转。

→ 思想：存在一个不能直接求导的似然函数，给定初始参数，找到与初始参数使似然函数与下界函数，在下界函数上求极值来更新参数，然后用更新后的参数为初始值再次进行如上操作，如图：



一般用于处理有隐含变量的模型，其似然函数为：

$$l(\theta)=\sum_{i=1}^{m}\log p(x^{(i)};\theta)=\sum_{i=1}^{m}\log\sum_{z^{(i)}} P(x^{(i)},z^{(i)};\theta)$$

→ 通过找下界巧妙的把难加的符号转移到log之外

其实EM的一般化形式中，所用的目标函数有为：

$$J(Q,\theta)=\sum_{i=1}^{m}\sum_{z^{(i)}} Q(z^{(i)})\log\frac{P(x^{(i)},z^{(i)};\theta)}{Q(z^{(i)})}$$，这样，EM算法就可以看做是对目标函数为坐标上升过程：E-Step中，θ不变，调整Q使函数变大；在M-Step中，Q不变，调整θ使函数变大。对 $l(\theta)$，对于当前 $\theta^{(t)}$，基在下界上通过极大似然估计得到 $\theta^{(t+1)}$，满足 $l(\theta^{(t+1)})\geq l(\theta^{(t)})$，即只有才能在下界函数上不断进行极大似然估计，逼近真实值。

$$l(\theta^{(t+1)})\geq \text{low\_bound}(\theta^{(t+1)})\geq \text{low\_bound}(\theta^{(t)})=l(\theta^{(t)})$$

MoG：混合高斯模型
MoMB：混合贝叶斯模型
因子分析模型
} → 与EM相关.

→ 要想用EM在这二模型上得到较好结果需满足 M(样本数) >> n(样本维度)。(这个问题类似于，线性方程组求解，但未知数个数比方程组数目多，因而不能完全求出未知数。)

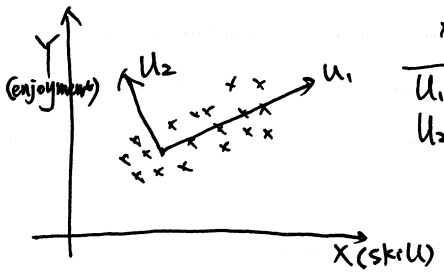
降维方法

→ 假设有隐含变量 $z \sim N(0,1)$ $z \in \mathbb{R}^d$ $(d<n)$
假设训练样本 X 由隐含变量 z 生成，即 $x = \mu + \Lambda z + \varepsilon$
其中：$\varepsilon \in N(0, \Psi)$。等价于，$z$ 已知时 x 的概率分布为：$x|z \sim N(\mu+\Lambda z, \Psi)$ 这就是因子分析模型的定义。其中：$\mu \in \mathbb{R}^n$，$\Lambda \in \mathbb{R}^{n \times d}$，$\Psi \in \mathbb{R}^{n \times n}$ 也是对角矩阵。

主成分分析(PCA)：将为余属性去掉；为方便计算需预处理 ⇒
1. $\mu = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$ 2. $x^{(i)} - \mu$ 替换 $x^{(i)}$
3. $\sigma_j^2 = \frac{1}{m}\sum_i (x_j^{(i)})^2$ 4. $x_j^{(i)} / \sigma_j$ 替换 $x_j^{(i)}$
1、2 使均值为0
3、4 使每个维度方差为1，即统一每个维度的尺度

寻找主方向，使点在主方向上的投影点方差最大化，
然后忽略掉主方向上的影响（降维），继续寻找主方向 u
最终得到 n 个主方向 u（相互正交）。选取方差最大的 k 个 u，则新的样本为（降维后）

$$y^{(i)} = \begin{bmatrix} u_1^T x^{(i)} \\ u_2^T x^{(i)} \\ u_3^T x^{(i)} \\ \vdots \\ u_k^T x^{(i)} \end{bmatrix}$$



x_1, x_2 有明显关联性。
u_1 反映了相关性
u_2 则是主方向之外的噪音

PCA 的一个应用：隐含语义索引 LSI
PCA 的一种实现：奇异值分解 SVD

	密度估计法(概率方法)	非概率方法
降维到子空间	因子分析	PCA
假设数据位于区块中	混合高斯模型	k-Means

No.

Date . .

有监督学习：告诉算法每个样本的正确答案，学习后的算法，对新输入也能输出正确答案。

1-8

1. 回归问题：需要预测的变量是连续的 → 用最小二乘法，做契合度评估
   → 求解方法：梯度下降 / 正规方程组
   $\theta=(X^TX)^{-1}X^T\vec{y}$
2. 分类问题：需要处理的问题是离散的

无监督学习：

9-11

根据样本集大小分为：

A. 过拟合（overfitting），特征项过大，
   使用 特征选择算法，一类自动化算法，选取用到的特征

B. 欠拟合（underfitting），特征项过小。
   使用 局部加权回归 来缓解对于选取特征的需求

学习（型）算法的理论基础「

非参数学习算法 / 参数学习算法：有固定数目参数的数据拟合算法

→ 参数数目会随训练集增长而变化的算法，可便于学习之后

又称为 loess.

强化学习：

1. 回报函数（正反馈/负反馈）

回归算法：logistic ⇒（分类算法，回归拟合）

$L(\theta)=P(\vec{y}|X;\theta)$ 「似然性

$l(\theta)=\log L(\theta)$ 对数似然性

使用梯度下降方法，使求似然性最大

$\theta_j:=\theta_j+\alpha\nabla_\theta(l(\theta))$

因为是求最大，所以梯度上升

偏导数计算后得到

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

$\nabla_\theta(\theta)=\frac{\partial}{\partial\theta_j}l(\theta)$

以上类似于梯度上升算法，形式上和线性回归是相同的，只是符号相反，但 $h_\theta(x^{(i)})$ 是 logistic 函数，和线性回归的实质是不一样的

采用 $g(z)=\frac{1}{1+e^{-z}}$ 为 sigmoid 函数

1. 每一次预测，都要重新拟合一条曲线
2. 权重 $w^{(i)}=\exp\left(-\frac{(x^{(i)}-x)^2}{2\tau^2}\right)$，即距离 $x^{(i)}$ 越近要预测的输入 x，$w^{(i)}$ 的值越接近 1。
3. τ 越宽波长函数，越小，钟型越窄
4. 若训练集很大，则用于预测的训练集就很大，工作量就可观了，此时要用 KD-tree 了

以上可以思考并简单解释：

1. 线性回归中，为什么选择最小二乘作为计算参数的指标，即假设预测出的值与真实值之间的面积的平方最小化？

2. 极大似然估计：选择 θ 使似然性 $L(\theta)$ 最大（数据出现的可能性最大），则要使 $\frac{1}{2}\sum_{i=1}^{m}(y^{(i)}-\theta^Tx^{(i)})^2$ 最小，即之前的 $J(\theta)$。所以可知：之前的最小二乘法，计算参数，实际上假设了误差项满足之前分布，且独立分布的情况，使似然最大化来计算参数。

牛顿方法，可以应用于 logistic 回归模型，加速拟合过程。

No.

Date

对于：$P(y|x;\theta)$

如果，y为实数，满足高斯分布，得到基于最小二乘法的线性回归

如果 y为$\{0,1\}$，满足伯努利分布，得到 Logistic 回归.

△选定了一个指数分布族后，怎么来推导出一个GLM（广义线性模型）呢？

No.

Date . .

生成学习算法

高斯判别分析 GDA

朴素贝叶斯

Laplace平滑.

sigmoid函数.

↳ 阈值函数

计算 $\theta^T x$，输出 $1 \Leftrightarrow \theta^T x \geq 0$，$0 \Leftrightarrow \theta^T x < 0$

对于如果 $\theta^T x \gg 0$，相当于确定的预测 $y=1$，

如果 $\theta^T x \ll 0$，相当于确定的预测 $y=0$

对于所有 i，如果 $y=1$，$\theta^T x^{(i)} \gg 0$，如果 $y=0$，$\theta^T x^{(i)} \ll 0$，则我们认为分类器是良好的

线性分类方法：Logistic回归 [通过牛顿方法或梯度下降得到一条直线，分割样本]

↑

指数分布族 [后验分布是 logistic 函数]

非线性分类方法：神经网络.

支持向量机 ⇒ 线性分类器. 可以拓展为非线性.

(函数间隔、几何间隔)?

超平面 (w,b) 和某特定训练样本 $(x^{(i)}, y^{(i)})$ 的函数间隔定义为：$\hat{\gamma}^{(i)} = y^{(i)}(w^T x + b)$

↳ 表示分类的准确度和可信程度

$wx+b=0$

↳ 意义 确定时：$|w \cdot x + b|$ 表示点到面的距离

↳ 法向量.

非凸性约束

$\max \frac{1}{\|w\|} \Rightarrow \min \frac{1}{2}\|w\|^2$

拉格朗日算子

拉格朗日函数

对偶问题

No.

Date

$ax+b+y=0$

$wx+b$