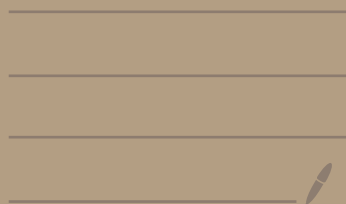


数据挖掘



(分类)

$h_{\text{logit}}(x)$ 输出的是一个标量

朴素贝叶斯, $y \in (0, 1)$ $x = (x_1, \dots, x_d)$ d 表示特征维数.
 $\arg\max_{\theta} f(\theta)$ 表示在 $f(\theta)$ 取最大值时, θ 的取值 $\arg\min$ 同理.

最优化理论与算法

① 梯度

$x \in \mathbb{R}^n$ 意味着 x 是一个向量 $x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$

② 凸集

凸组合

设 $S \subseteq \mathbb{R}^n$, $\forall x^{(1)}, x^{(2)} \in S$, $\forall \lambda \in [0, 1]$, 有 $\lambda x^{(1)} + (1-\lambda)x^{(2)} \in S$, 则得 S 为凸集.

n 维空间中的子集

两个点(向量)

两个点连线上的某一点, 表示出来的其实是一个点集.

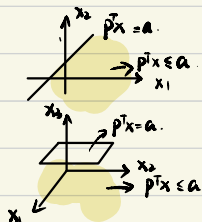
半空间 $H = \{x | p^T x \leq a\}$

超平面

$(p_1, p_2, \dots, p_n) \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = a$
常数

① 当 $n=2$, $\{x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} | (p_1, p_2) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \leq a\}$

② 当 $n=3$, $\{x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} | (p_1, p_2, p_3) \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \leq a\}$



③ 证对 $\forall n$, $H = \{x | p^T x \leq a\}$ 是凸的

$\forall x^{(1)}, x^{(2)} \in H$, 满足 $p^T x^{(1)} \leq a$, $p^T x^{(2)} \leq a$.

则 $\forall \lambda \in [0, 1]$, $p^T(\lambda x^{(1)} + (1-\lambda)x^{(2)}) = \lambda p^T x^{(1)} + (1-\lambda)p^T x^{(2)} \leq \lambda a + (1-\lambda)a = a$.

$\therefore \lambda x^{(1)} + (1-\lambda)x^{(2)} \in H$ $\therefore H$ 为凸集

凸集的极点 (不能在凸集中两个不同的点, 使其在这两个点之间)

范数 (Norm): 用于刻画距离

① $\forall x \in \mathbb{R}^n$ $\|x\| \geq 0$

$\|x\| = 0$, 当且仅当 (iff) $x = 0$

② $\|ax\| = |a| \cdot \|x\|$

③ $\|x+y\| \leq \|x\| + \|y\|$

设 $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$

L_1 范数 $\|x\|_1 = \sum |x_i| = n$

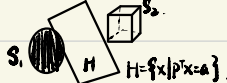
L_2 范数 $\|x\|_2 = \sqrt{x_1^2 + \dots + x_n^2} = \sqrt{n}$

L_∞ 范数 $\|x\|_\infty = \max_j |x_j|$

L_p 范数 $\|x\|_p = (\sum |x_j|^p)^{\frac{1}{p}}$

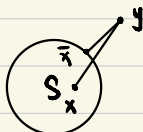
凸集分离定理

超平面 H 分离凸集 S_1 和 S_2 , 如果对 $\forall x \in S_1$, 有 $p^T x \geq a$, $\forall x \in S_2$, 有 $p^T x \leq a$, 则称 H 分离集合 S_1, S_2 .



定理:

二维:



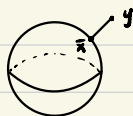
y 是凸集外的点, $y \notin S$ → 闭凸集

$x \in S$ 唯一 $\exists \bar{x} \in S$, 使得 $\|y - \bar{x}\| = \inf_{x \in S} \|y - x\|$

若 S 为非凸, 则 $\bar{x}$ 则不唯一.

若 S 非闭, 则找不到与 y 最近.

三维:



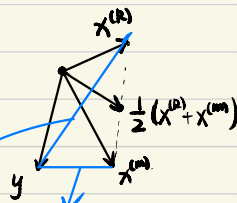
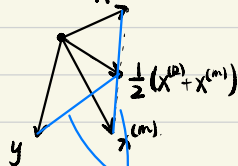
证明

(可以当作取最小值)

向量序列

设 $r = \inf_{x \in S} \|y - x\| > 0$, 由下确界定义, 存在序列 $\{x^{(k)}\} \in S$, 使得 $\|y - x^{(k)}\| \rightarrow r$

下证 $\{x^{(k)}\}$ 的极限点不存在, 且 $\bar{x} \in S$.



由平行四边形律

$$\begin{aligned} \text{则 } \|x^{(k)} - x^{(m)}\|^2 + 2\left\|\frac{x^{(k)} + x^{(m)}}{2} - y\right\|^2 &= 2\|x^{(k)} - y\|^2 + 2\|x^{(m)} - y\|^2 \\ \|x^{(k)} - x^{(m)}\|^2 + 4\left\|\frac{x^{(k)} + x^{(m)}}{2} - y\right\|^2 &= 2\|x^{(k)} - y\|^2 + 2\|x^{(m)} - y\|^2 \\ &\geq 4r^2 \end{aligned}$$

$$\text{则 } \|x^{(k)} - x^{(m)}\|^2 \leq 2\|x^{(k)} - y\|^2 + 2\|x^{(m)} - y\|^2 - 4r^2$$

第一步, 存在性

当 k, m 足够大时 $\|x^{(k)} - x^{(m)}\| \rightarrow 0$, 故 $\{x^{(k)}\}$ 为 Cauchy 序列.

则 $\{x^{(k)}\}$ 极限存在, 设为 $\bar{x}$ $\because S$ 为闭集 $\therefore \bar{x} \in S$.

假设 $\exists \hat{x} \in S$, 使 $\|y - \hat{x}\| = \|y - \bar{x}\| = r$ $\because S$ 为凸集 $\therefore \bar{x}, \hat{x} \in S \therefore \frac{\bar{x} + \hat{x}}{2} \in S$

$$\therefore r \leq \|y - \frac{\bar{x} + \hat{x}}{2}\| \leq \frac{1}{2}\|y - \bar{x}\| + \frac{1}{2}\|y - \hat{x}\| = r \Rightarrow r \leq \|y - \frac{\bar{x} + \hat{x}}{2}\| \leq r$$

拆分, 两边取范数可得.

$$\text{则 } \|y - \frac{\bar{x} + \hat{x}}{2}\| = r$$

由平行四边形律:

$$\|\hat{x} - \bar{x}\|^2 + 4\left\|y - \frac{\bar{x} + \hat{x}}{2}\right\|^2 = 2\|y - \hat{x}\|^2 + 2\|y - \bar{x}\|^2$$

$$\|\hat{x} - \bar{x}\|^2 + 4r^2 = 2r^2 + 2r^2 \therefore \|\hat{x} - \bar{x}\|^2 = 0$$

第二步, 唯一性

则 $\hat{x} = \bar{x}$, 则 $\bar{x}$ 唯一

(回归)

线性回归

$h_{\text{regression}}(x)$ 输出的是一个回归值

$y_i \in \mathbb{R}$ (y_i 为回归值)

$x_i = (x_{i1}, \dots, x_{ip})^T, i=1, 2, \dots, n$ (i 是对 data 计数, p 是特征维度)

$x \in \mathbb{R}^p$ $w^T x = x^T w = \langle w, x \rangle = \sum_{j=1}^p w_j x_j = f(x, w)$ 要求的

因为 i 留给 data 计数 来自 data sample



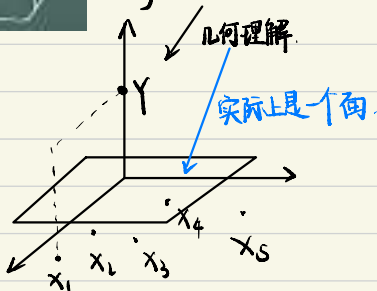
data sample:

$$x_1 = (x_{11}, x_{12})^T \quad y_1$$

$$x_n = (x_{n1}, x_{n2})^T \quad y_n$$

挑选线性模型

$$z = f(x) = w^T x = w_1 x_1 + w_2 x_2$$



最小二乘线性回归 $\rightarrow$ 对正值进行处理

$$J(w) = \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i, w))^2$$

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix} \quad Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad Y - Xw = \begin{pmatrix} y_1 - x_1^T w \\ \vdots \\ y_n - x_n^T w \end{pmatrix}$$

$$\text{则 } \|Y - Xw\|_2^2 = \sum_{i=1}^n (y_i - f(x_i, w))^2$$

$$\therefore J(w) = \frac{1}{n} \|Y - Xw\|_2^2 \rightarrow \text{判断拟合好坏 (优化目标)}$$

$$= \frac{1}{n} (Y - Xw)^T (Y - Xw) \quad \text{其中 } w \text{ 为变量}$$

目标: 求 $J(w)$ 最小时的 w $w^* = \arg\min_{w} J(w)$

计算梯度: $0 = \frac{\partial J(w)}{\partial w} = \frac{\partial (\frac{1}{n} (Y - Xw)^T (Y - Xw))}{\partial w} = \frac{1}{n} \cdot \frac{\partial (Y^T - w^T X^T) (Y - Xw)}{\partial w}$

$$= \frac{1}{n} \frac{\partial (Y^T Y - Y^T X w - w^T X^T Y + w^T X^T X w)}{\partial w}$$
$$= \frac{1}{n} (0 - X^T Y - X^T Y + 2 X^T X w)$$
$$= \frac{1}{n} (-X^T Y + X^T X w)$$

得 $\hat{w} = (X^T X)^{-1} X^T Y$

显式解, 缺点, 不一定算得出来, 因为可能不是可解

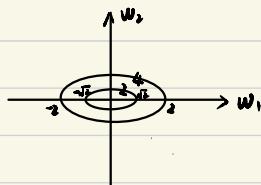
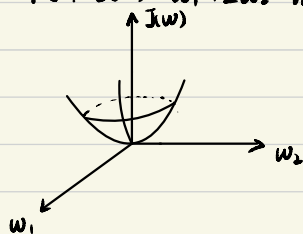
当 x 比如 $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ 时不可逆

$$\frac{\partial Y^T X w}{\partial w} = X^T Y$$

原理: $\frac{\partial \beta^T w}{\partial w} = \frac{\partial (b_1 w_1 + \dots + b_n w_n)}{\partial w} = \beta$

也可记: $\nabla \beta^T x = \beta$

$w \in \mathbb{R}^2$. $J(w) = w_1^2 + 2w_2^2$ 请绘制 $J(w)$ 空间曲线图和等高线图.



$f(w) = w_1^2 + w_2^2$. 若看到左式, $w \in \mathbb{R}^2$ (w 是几维向量)

w 可以是任意维度, 只是 $f(w)$ 取的是 w 前2维.

Loss 函数: $L = J(w) - \text{正则项}$ $\rightarrow$ 优化目标.

梯度下降法: $w^{(0)} \in \mathbb{R}^p$ // 初始化

$$w^{(n+1)} = w^{(n)} - \eta \frac{\partial}{\partial w} (J(w)) \quad n=0, 1, 2 \dots \text{直到收敛}$$

学习率 (快慢调节)

岭回归 $\min_w \frac{1}{n} \|y - Xw\|_2^2 \quad \text{s.t. } \|w\|_2 < C$

常数, w 取值范围



Lasso 回归 $\min_w \frac{1}{n} \|y - Xw\|_2^2 \quad \text{s.t. } \|w\|_1 < C$



最优性条件

[F-J] 条件: $\bar{w}$ 是局部最优解, $\exists$ 不全为零的非负数 $\alpha_1, \alpha_2, \dots$, 使得 $\alpha_1 \nabla J(\bar{w}) - \sum \alpha_i \nabla g_i(\bar{w}) = 0$

[KT 条件]: 若 $\{\nabla g_i(w)\}$ 线性无关, 再若 $\bar{w}$ 是局部最优解, $\exists$ 不全为零的 $\alpha_1, \alpha_2, \dots$

应用

$$\text{使得 } \nabla J(\bar{w}) = \sum \alpha_i \nabla g_i(\bar{w})$$

Ridge (岭回归) 和 Lasso.

正则项.

Loss 函数可写为 $\frac{1}{n} \|y - Xw\|_2^2 + w(C \|w\|_2 - C)$

线性回归

岭回归

线性回归的基础上加入正则项, 避免过拟合

LASSO 回归

$$J = \frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2$$

$$= \frac{1}{n} \sum_{i=1}^n (w^T x_i + b - y_i)^2$$

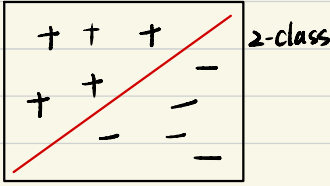
均方误差, 它衡量模型预测值与实际值之间的差异

$$J = \frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2 + \lambda \|w\|_2^2$$

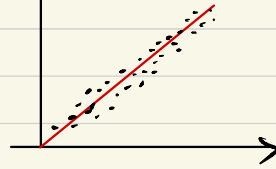
正则项实际的作用是模型参数越小越好, 也就是让, 希望忽略一些不重要的特征.

$$J = \frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2 + \lambda \|w\|_1$$

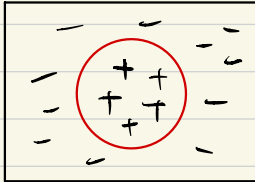
① linear - classification



③ linear - Regression



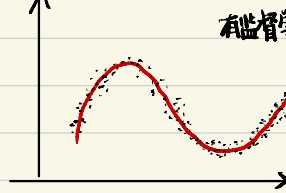
②



有监督学习 - label

nonlinear - classification

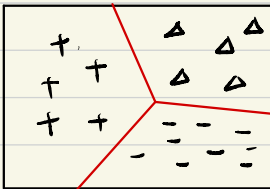
④



有监督学习 - Regression value

nonlinear - Regression

⑤



multi-class

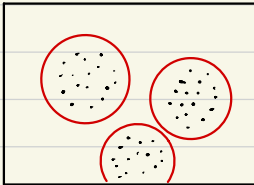
classification

⑦



描述一个曲面 contour

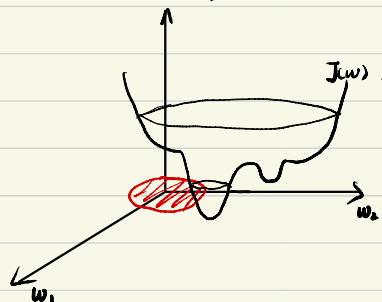
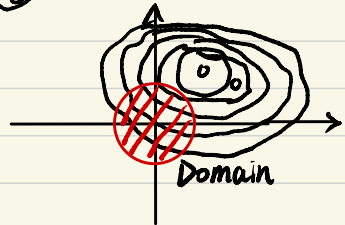
⑥



unsupervised-learning
无监督

cluster (无标签 label)

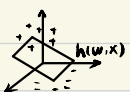
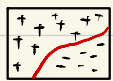
⑧



在凸函数中, 局部最小解就是全局最小解.

支持向量机

① 超平面 $w^T x + b = 0$ $= h(w, x)$

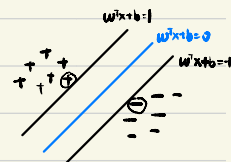
② 场景:  非线性场景 

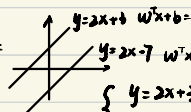
③ 符号准备: data sample (x^i, y^i) $i=1, 2, \dots$ (对样本计数)
 $x^i \in \mathbb{R}^d$ (d 维) $y^i = 1$ 或 -1

④ 超平面分类器的建模

1' 目标: 把 $+$ / $-$ 数据截然分开

$w^T x + b > 0$ $w^T x + b = 0$ $w^T x + b < 0$ 求 w, b .
 $y^i = 1$ $y^i = -1$
 约束条件 $y^i (w^T x^i + b) \geq 1$



eg:  $x = \begin{pmatrix} x \\ y \end{pmatrix}$

$$\begin{cases} y = 2x + 3 \rightarrow -2x + y = 3 \rightarrow -2x + y + 2 = 5 \\ y = 2x - 7 \rightarrow -2x + y = -7 \rightarrow -2x + y + 2 = -5 \end{cases}$$

$$\begin{cases} -\frac{2}{5}x + \frac{1}{5}y + \frac{2}{5} = 1 \rightarrow (-\frac{2}{5}, \frac{1}{5}) \begin{pmatrix} x \\ y \end{pmatrix} + \frac{2}{5} = 1 \\ -\frac{2}{5}x + \frac{1}{5}y + \frac{2}{5} = -1 \rightarrow (-\frac{2}{5}, \frac{1}{5}) \begin{pmatrix} x \\ y \end{pmatrix} + \frac{2}{5} = -1 \end{cases}$$

max margin

2' 分得最合理 (找最宽的间隔)

如何计算间隔宽度? (点到面的距离公式) 点 x 到 $w^T x + b = 0$ 的距离.

$$d(x, w, b) = \frac{|w^T x + b|}{\|w\|_2} \xrightarrow{\text{二维样本 } x \in \mathbb{R}^2} x = \begin{pmatrix} x \\ y \end{pmatrix}$$

点 $\rightarrow$ 线距离: 点 (x_0, y_0) , 平面 xoy , 直线 $l: w_1 x + w_2 y + b = 0$.

$$\text{则 } (x_0, y_0) \text{ 到 } l \text{ 距离为: } d = \frac{|w_1 x_0 + w_2 y_0 + b|}{\sqrt{w_1^2 + w_2^2}}$$

$$\text{点} \rightarrow \text{面距离: 点 } (x_0, y_0, z_0), \text{面 } Ax + By + Cz + D = 0 \quad d = \frac{|Ax_0 + By_0 + Cz_0 + D|}{\sqrt{A^2 + B^2 + C^2}}$$

④.1 最优分类面的 margin 计算

$$x^i \in \mathbb{R}^d \text{ 到 } w^T x + b = 0 \quad d = \frac{|w^T x + b|}{\|w\|_2}, \text{ 记 } d(x^i, w, b)$$

$$\text{margin 宽度: } \min_{x^i} d(x^i, w, b) + \min_{y^i} d(x^i, w, b) = \frac{|-1|}{\|w\|_2} + \frac{|1|}{\|w\|_2} = \frac{2}{\|w\|_2}$$

④② $\max \text{margin} \Leftrightarrow \min \phi(w) = \frac{\|w\|^2}{2}$

④③ 在线性可分的情形下, 优化目标为 $\min \phi(w) = \frac{\|w\|^2}{2} \quad \text{s.t. } y^i(w^T x^i + b) \geq 1$
类似 $\min J(w) \quad \text{s.t. } \|w\| < C$

④④ 用对偶问题来表征和求解上述模型.

$\phi(w, b, d) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^n d_i [y^i(w^T x^i + b) - 1]$

对偶问题: $\max_w \phi(w) = \max_{w, b} (\min_{d_i} \phi(w, b, d))$

④⑤ 求 $\min_{w, b} \phi(w, b, d) \rightarrow \begin{cases} \nabla_w \phi(w, b, d) = 0 \\ \frac{d}{dw} \phi(w, b, d) = 0 \end{cases} \quad (0)$

解: $\nabla \|w\|^2 = \nabla (w^T w) = 2w \quad \leftarrow \begin{matrix} w \in \mathbb{R}^d \\ \|w\|^2 = w_1^2 + \dots + w_d^2 = w^T w \end{matrix}$

$\nabla w(w^T x^i + b) = x^i + 0 = x^i$

$\therefore \nabla_w \phi(w, b, d) = 0 \quad \text{即} \quad \nabla_w (\frac{1}{2} \|w\|^2 - \sum_{i=1}^n d_i (y^i(w^T x^i + b) - 1)) - w - \sum_{i=1}^n d_i y^i x^i = 0$

则 $w^* = \sum_{i=1}^n d_i y^i x^i \quad (\text{记 } w^*, b^* = \min_{w, b} \phi(w, b, d))$

易知 $\frac{d}{db} \phi(w, b, d) = -\sum_{i=1}^n d_i y^i = 0$

④⑥ $\max_{d_i} \phi(w^*, b^*, d) = \max_{d_i} \phi(w^*, b^*, d)$
 $= \max_{d_i} (\frac{1}{2} \|w^*\|^2 - \sum_{i=1}^n d_i (y^i(w^{*T} x^i + b) - 1))$
 $= \max_{d_i} (\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n d_i d_j y^i y^j x^{iT} x^j - \sum_{i=1}^n d_i (y^i (\sum_{j=1}^n d_j y^j x^{iT}) x^i + b) - 1)$
 $= \max_{d_i} \{ \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n d_i d_j y^i y^j \langle x^i, x^j \rangle - \sum_{i=1}^n d_i y^i \langle x^i, x^i \rangle - \sum_{i=1}^n d_i y^i b + \sum_{i=1}^n d_i \}$
 $= \max_{d_i} \{ -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n d_i d_j y^i y^j \langle x^i, x^j \rangle + \sum_{i=1}^n d_i \}$
 d^* 的求解可写为 $d^* = \min_{d_i} \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n d_i d_j y^i y^j \langle x^i, x^j \rangle - \sum_{i=1}^n d_i$

内外两圈循环
编程易于实现

5. SVM classifier 超平面分类 (有噪音) 线性可分到不可分的跨越

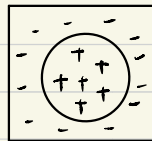
⑤① 容错处理: $y^i(w^T x^i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0$
修正: $\min_{w, \xi} \phi(w, \xi) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i$

s.t. $\begin{cases} y^i(w^T x^i + b) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$

往下求解与④中类似

6. 不可分场景 (核方法)

$$\phi: \begin{matrix} \text{低维} \rightarrow \text{高维} \\ R^2 \rightarrow R^3 \\ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} x_1 \\ x_2 \\ x_1^2 + x_2^2 \end{pmatrix} \end{matrix}$$



eg. $w^T x + b = (1, 1) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + b = x_1 + x_2 + b = 0$

↓ 高维.

$$x_1^2 + x_1 x_2 + x_2^2 + b = 0 \longrightarrow (x_1 - \frac{1}{2}x_2)^2 + \frac{\sqrt{3}}{2}x_2^2 + b = 0$$

PCA 主成分分析

若干结论准备:

1. $d, \beta \in R^2$, d 向 β 投影, 求投影向量表达式.

R^2 记 r 为 d 向 β 作投影

$$\langle d, \beta \rangle = |d| \cdot |\beta| \cos(\angle d, \beta)$$

$$r = \frac{|d| |\beta| \cos(\angle d, \beta)}{|\beta| |\beta|} \cdot \beta = |d| \cos(\angle d, \beta) \cdot \frac{\beta}{|\beta|} \quad (\beta \text{ 方向单位向量})$$

$$\longrightarrow r = \frac{\langle d, \beta \rangle}{\langle \beta, \beta \rangle} \cdot \beta$$

2 (多元统计知识) $x = \begin{pmatrix} x_1 \\ x_2 \\ x_n \end{pmatrix}$ 为 n 元随机变量, $\mu = E[x] = \begin{pmatrix} E[x_1] \\ E[x_2] \\ E[x_n] \end{pmatrix}$ 为 x 的期望

$\Sigma = (\Sigma_{ij})_{n \times n}$ 为协方差矩阵, $\Sigma_{ij} = E[(x_i - \mu_i)(x_j - \mu_j)]$

问: $\Sigma \stackrel{?}{=} E[(x - \mu)^T (x - \mu)]$ 或 $\Sigma \stackrel{?}{=} E[(x - \mu)(x - \mu)^T]$ 哪个结论正确?

x 是个列向量 $(x - \mu)_{n \times 1}$, 则 $(x - \mu)^T_{1 \times n}$ 才可以形成 $n \times n$ 矩阵

当 $n=1$ 时 (1 维) 方差 $\text{Var} = \sigma^2 = E[(x - E(x))^2]$

3. 设上述 x 的观测值集合 $O = \{ \overset{\text{observation}}{x^{(1)}}, x^{(2)}, \dots, x^{(m)} \}$ 问记号 x_i 与 $x^{(i)}$ 的区别.

x_i : 随机变量 $O = \{ x^{(1)} = \begin{pmatrix} x_1 \\ x_2 \\ x_n \end{pmatrix}, \dots, x^{(m)} = \begin{pmatrix} x_1 \\ x_2 \\ x_n \end{pmatrix} \}$

$x^{(i)} = \begin{pmatrix} x_1^{(i)} \\ x_2^{(i)} \\ x_n^{(i)} \end{pmatrix}$

4. $\hat{\mu} = \frac{1}{m} \sum_{j=1}^m x_j^{(1)}$, $\hat{\Sigma} = \frac{1}{m} \sum_{j=1}^m (x_j - \mu)(x_j - \mu)^T$ 问 $\hat{\mu}$ 和 $\hat{\Sigma}$ 是什么? 它们与 μ 和 Σ 的关系是什么?

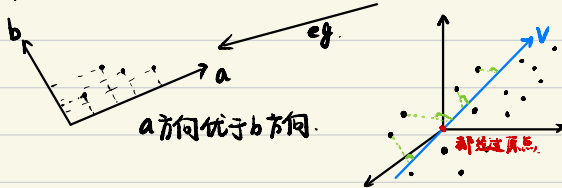
$\hat{\mu}$ 是 μ 的估计, 叫最大似然估计, $\hat{\Sigma}$ 根据样本可以求协方差的估计

5. $C_{n \times n}$ 为对称矩阵, $X \in \mathbb{R}^{n \times 1}$

$$\max_{X \in \mathbb{R}^{n \times 1}} X^T C X \quad \text{s.t.} \quad \|X\|_2 = 1 \quad \text{该优化问题如何求解?}$$

PCA 希望在 n 维空间中找到若干条最好的线/方向, 向上作投影, 以线/方向作新的坐标轴, 并用新的坐标代替原来的.

希望新的投影散的开



$X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ 为 n -元随机变量.

$O = \{x^{(1)}, x^{(2)}, \dots, x^{(n)}\}$ 为观测向量集合.

每个点可表示为 O 中的一个 $x^{(i)}$.

V 为某投影方向向量.

$$x^{(i)} \text{ 向 } V \text{ 投影, 投影向量: } \frac{\langle x^{(i)}, V \rangle}{\langle V, V \rangle} \cdot V = \frac{\langle x^{(i)}, V \rangle}{\|V\|^2} V$$

$$\text{不妨设 } \|V\| = 1 \quad (V^T V = \langle V, V \rangle = 1, \|V\| = 1) \quad x^{(i)T} \cdot V = V^T \cdot x^{(i)}$$

$$\begin{aligned} \sigma^2 \text{ 为所有投影坐标的方差, 则 } \sigma^2 &= \frac{1}{n} \sum_{i=1}^n (V^T x^{(i)} - \bar{o})^2 = \frac{1}{n} \sum_{i=1}^n (V^T x^{(i)}) (V^T x^{(i)})^T \\ &= \frac{1}{n} \sum_{i=1}^n V^T x^{(i)} x^{(i)T} V = V^T \left(\frac{1}{n} \sum_{i=1}^n x^{(i)} x^{(i)T} \right) V \\ &= V^T C V \end{aligned}$$

则 PCA 优化问题即: $V = \operatorname{argmax}_{V \in \mathbb{R}^n, \|V\|=1} V^T C V$ 拉格朗日乘 $f(V, \lambda) = V^T C V - \lambda(V^T V - 1)$

梯度求解 $\rightarrow \frac{\partial f}{\partial V} = 0$, 即 $\frac{\partial f}{\partial V} = 2CV - 2\lambda V = 0$ 求 $\hat{V}$, $C\hat{V} = \lambda\hat{V}$.

下
对称矩阵

PCA 最后的问题就是对协方差矩阵求特征值, 特征向量.

最后求得有若干条 V , 可以按特征值绝对值进行排序 $(V_1, V_2, \dots)$, V_1 重要性最高.

K-Means:

K-means算法是一种常见的聚类算法。它的主要目标是将数据分为K个不同的簇。以下是K-means算法的基本计算流程：

1. 初始化：选择K个数据点作为初始的簇中心。
2. 分配：将每个点分配到最近的簇中心，形成K个簇。
3. 更新：重新计算每个簇的中心，通常是簇内所有点的均值。
4. 迭代：重复步骤2和步骤3，直到簇中心不再变化或达到预定的迭代次数。
5. 终止：当簇中心在连续几次迭代中的移动距离小于某个阈值，或者达到预定的迭代次数时，算法终止。

具体步骤如下：

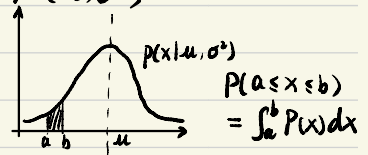
1. 选择初始簇中心：可以随机选择，也可以使用如K-means++等更高级的方法来选择初始簇中心。
2. 为每个点分配簇：对于每个数据点，计算它到每个簇中心的距离，并将该点分配到最近的簇中心。
3. 更新簇中心：对每个簇，计算所有分配到该簇的点的均值，并将这个均值作为新的簇中心。
4. 检查收敛：比较新的簇中心和旧的簇中心，如果簇中心没有显著变化（例如，移动距离小于某个阈值），则算法可以停止。否则，返回步骤2继续迭代。
5. 结束：当满足终止条件时，算法结束。最终的簇中心和簇分配即为算法的输出。

单高斯模型

Random Variable (R.V), $X \sim N(\mu, \sigma^2)$

PDF (概率密度函数): $p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2}(x-\mu)^2)$

or $p(x|\mu, \sigma^2)$



概率模型的 General Paradigm:

- 1) 假定数据服从某分布 (eg. Gaussian).
- 2) 抽样 估计分布的参数 (μ, σ^2)

最大似然估计 (MLE) 求解单高斯模型

1) 构造似然函数: $L(\mu, \sigma^2) = \prod_{i=1}^N p(x_i | \mu, \sigma^2)$

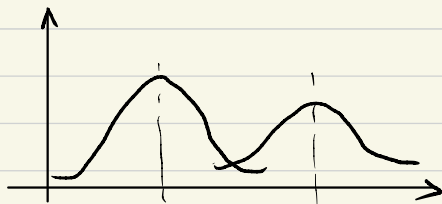
2) 取对数: $l(\mu, \sigma^2) = \log L(\mu, \sigma^2) = \sum_{i=1}^N \log p(x_i | \mu, \sigma^2)$

3) 最大似然估计: $\log$ 单调 $\uparrow$, 求导 $= 0$, 找到 $\hat{\mu}$ 和 $\hat{\sigma}^2$

$$\frac{\partial l}{\partial \mu} = 0 \Rightarrow \hat{\mu} = \frac{1}{N} \sum x^{(i)} \quad \hat{\sigma}^2 = \frac{1}{N} \sum (x^{(i)} - \hat{\mu})^2$$

第1个隐变量模型

高斯混合模型 (Gaussian mixture model)



(μ, σ^2) 混合系数

通过调整每个高斯分布的参数，
使得所有数据点联合概率最大。

GMM与K-Means的联系与区别

联系：

- 1) GMM和K-Means都是将数据分成多个簇。
- 2) 都通过迭代优化来优化某个目标函数。

K-Means：簇内点到中心的距离之和。

GMM：对数似然函数。

区别：

- 1) K-Means：硬聚类；GMM：软聚类。
- 2) K-Means 参数少；GMM 参数更多，包括每个高斯分布均值、协方差和权重。

混合高斯模型的EM算法求解

所有观测值放入 $\bar{x} = \{x^{(1)}, x^{(2)}, \dots, x^{(n)}\}$ 。每个样本独立同分布且符合高斯分布

K个高斯分布 $N(\mu_k, \sigma_k^2)$

$\theta = \{\mu_1, \sigma_1^2, \dots, \mu_K, \sigma_K^2\}$

△引入权重 α_k ，定义 likelihood $P(x|\theta)$

$$p(x|\theta) = \sum \alpha_k N(x; \mu_k, \sigma_k^2) \text{ 且使 } \sum \alpha_k = 1$$

α_k 应与 x 有关，不应是固定常值。

△引入隐变量 z

对每个 x 来说，都对应一个 z ， z 的取值为 $1, \dots, K$ ，即每个 $x^{(i)}$ 对应一个 $z^{(i)}$

* $\int$ $P(z|\theta)$ 替代 α_k

$P(z=k|\theta)$ ： z 抽中是 k 的概率。（随便一个点，这个点是蓝/黄点的概率）

$$p(x|z, \theta) = p(x|\mu_k, \sigma_k^2)$$

$p(x|\theta) = \prod_{i=1}^n p(x^{(i)}|\theta)$ 似然函数.
 $p(x|\theta) = \int p(x, z|\theta) dz = \int p(z|\theta) p(x|z, \theta) dz$ 边缘概率 (连续版本)
 $p(x|\theta) = \sum_{z=1}^K p(z|\theta) p(x|z, \theta)$ 离散版本.

E-M 步骤:

E步: 估计 z , 估计隐变量分布 $p(z|x, \theta)$

M步: Maximize, 最大化联合分布期望

$$\int \log p(x, z|\theta) p(z|x, \theta) dz = \arg \max_{\theta} E_{z \sim p(z|x, \theta)} [\log p(x, z|\theta)]$$

求解步骤:

① 似然函数 $\log p(x|\theta) = \log p(x, z|\theta) - \log p(z|x, \theta)$ $p(x) = \frac{p(x, z)}{p(z|x)}$

$$= E_{p(z|x, \theta^{(q)})} [\log p(x|\theta)]$$

$$= E_{p(z|x, \theta^{(q)})} [\log p(x, z|\theta) - \log p(z|x, \theta)]$$

$$\int \log p(x|\theta) \cdot p(z|x, \theta^{(q)}) dz = \underbrace{\int \log p(x, z|\theta) \cdot p(z|x, \theta^{(q)}) dz}_{Q(\theta, \theta^{(q)})} - \underbrace{\int \log p(z|x, \theta) \cdot p(z|x, \theta^{(q)}) dz}_{H(\theta, \theta^{(q)})}$$

定理: $H(\theta^{(q)}, \theta^{(q)}) \geq H(\theta, \theta^{(q)})$

$$\theta^{(q+1)} = \arg \max_{\theta} \int \log p(x, z|\theta) \cdot p(z|x, \theta^{(q)}) dz$$

$$= \arg \max_{\theta} Q(\theta, \theta^{(q)}) = \arg \max_{\theta} (\log p(x|\theta) + H(\theta, \theta^{(q)}))$$

$$\rightarrow \log p(x|\theta^{(q+1)}) + H(\theta^{(q+1)}, \theta^{(q)}) \geq \log p(x|\theta^{(q)}) + H(\theta^{(q)}, \theta^{(q)})$$

$$\therefore \log(p(x|\theta^{(q+1)})) - \log(p(x|\theta^{(q)})) \geq H(\theta^{(q+1)}, \theta^{(q)}) - H(\theta^{(q)}, \theta^{(q)}) \geq 0$$

$\theta^{(q+1)}$ 是一个正确的取最大化似然方法.

$$\begin{aligned}
 \sigma_{z^i}^{(q+1)} &= \frac{1}{N} \sum_{i=1}^N p(z^i|x^i, \theta^{(q)}) \\
 \mu_{z^i}^{(q+1)} &= \frac{\sum_{i=1}^N x^i p(z^i|x^i, \theta^{(q)})}{\sum_{i=1}^N p(z^i|x^i, \theta^{(q)})} \\
 \sigma_{z^i}^{(q+1)} &= \frac{\sum_{i=1}^N [x^i - \mu_{z^i}^{(q+1)}][x^i - \mu_{z^i}^{(q+1)}]^T p(z^i|x^i, \theta^{(q)})}{\sum_{i=1}^N p(z^i|x^i, \theta^{(q)})}
 \end{aligned}$$

第2个隐变量模型

LDA主题模型 (变分推断算法 (EM扩展))

(生成模型)

建立观测值与隐变量之间的分布依赖关系 —— 概率图模型

θ_d : 第d个文档主题的抽样 β_k : 第k个主题的抽样 (某主题下单词的概率)

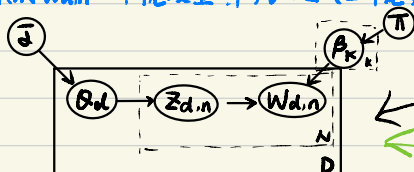
$w_{d,n}$: 第d个文档下第n个单词

$z_{d,n}$: 辅助变量, θ_d 中具体是哪一个主题取出来 $\rightarrow$ 知道具体的k.

$z_{d,n} \sim \text{Multinomial}(\theta_d)$

$w_{d,n} \sim \text{Multinomial}(\beta_{z_{d,n}})$

观测值只有 $w_{d,n}$, 隐变量: θ, β, z (三个隐变量类别)



VI (variational Inference) 变分推断

E步: 固定住 ELBO 中的模型参数, 优化变分参数, 刻画 $q(z)$ 变分函数

M步: 固定住 ELBO 中的变分函数, 优化模型参数, 刻画 θ 和 β 的概率图 (joint 分布)

E: $q(\theta_d), q(z_{d,n}), q(\beta_k) \rightarrow q(z)$

θ — 模型参数

$$\log p(w|\theta) = \log \int_{\theta} \int_{\beta} \sum_z p(\theta, z, \beta, w|\theta, \pi) d\theta d\beta \quad \text{引入隐变量}$$

$$= \log \int_{\theta} \int_{\beta} \sum_z \frac{p(\theta, z, \beta, w|\theta, \pi) \cdot q(\theta, z, \beta)}{q(\theta, z, \beta)} d\theta d\beta \quad (\text{变分推断})$$

$$\text{要极大化 } \log p(w|\theta) = \log E_{q(\theta, z, \beta)} \left[\frac{p(\theta, z, \beta, w|\theta, \pi)}{q(\theta, z, \beta)} \right] \geq E_{q(\theta, z, \beta)} \left[\log \frac{p(\theta, z, \beta, w|\theta, \pi)}{q(\theta, z, \beta)} \right]$$

可以通过极大化

ELBO 来完成.

$$= E_q [\log p(\theta, z, \beta, w|\theta, \pi)] - E_q [q(\theta, z, \beta)] \quad \text{证据下界 ELBO}$$

$$= p(\theta|\theta) p(z|\theta) p(\beta|\pi) p(w|z, \beta)$$

$$= E_q(\theta) [p(\theta|\theta)] + E_{q(z)} [p(z|\theta)] + E_{q(\beta)} [p(\beta|\pi)] + E_{q(z, \beta)} [p(w|z, \beta)] - E_{q(z)} [q(z)] - E_{q(\beta)} [q(\beta)] - E_{q(z, \beta)} [q(z, \beta)]$$

(Auto-Encoder)

VAE/变分自编码器

$$X = \{x_1, x_2, \dots, x_n\}$$

可以是一个或多个随机变量

① 生成模型：对 $P(X)$ 建模， X - observation

生成新 $\hat{x}$ ，逼近原 x

判别模型：对 $P(Y|X)$ 建模， X - observation， Y - 模型参数/数据标签(与 x 有关)

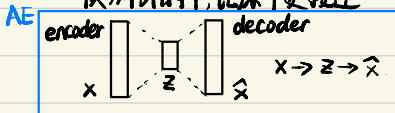
(SVM, NB, BP)

区别

预测出的 $\hat{y}$ ，比原 y 更接近 $\hat{y} \rightarrow y$

② AE 是不是隐变量模型？

是



③ VAE 与 AE 的区别，Variational (变分) 如何体现“变分方法”的特点？

隐变量 $z \sim q(z)$ 一个分布族

看哪个模型参数取下来 $q^*(z)$ 最好

概率密度函数 (PDF)

$\left\{ \begin{array}{l} \text{函数 } f(x) : x \text{ 取不同值, 结果不同} \\ \text{变分 } f_i(x) : f_i \text{ 不同, 结果不同} \end{array} \right.$

AE: 用多层神经网络训练一个小层, 再用小层重建

优化目标: $q(z) \rightarrow p(z|x)$

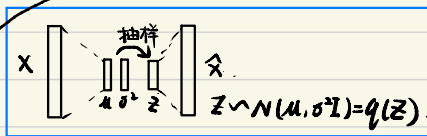
交叉熵

VAE 表达式 (loss)

$$\log p(x) - \text{KL}[q(z|x) \| p(z|x)] = \mathbb{E}_{z \sim q} [\log p(x|z)] - \text{KL}[q(z|x) \| p(z)]$$

证据概率(最大)

最小(希望 $q(z) \rightarrow p(z|x)$)



VAE

$$p(x) = \int p(x|z; \theta) p(z) dz$$

$$= \mathbb{E}_{z \sim q} [\log q(z|x) - \log p(z|x)]$$

$$= \mathbb{E}_{z \sim q} [\log q(z|x) - \log p(x|z) - \log p(z)] + \log p(x)$$

$$\frac{p(x|z) - p(z)}{p(x)}$$

则可得上述 VAE 表达式

重参 (Re-Parameter)

$$z = \mu(x) + \Sigma^{\frac{1}{2}}(x) \cdot \epsilon$$

使抽样不被打断