Section 4
Linear Classification

① 背景

② 感知机

③ 线性判别分析

④ 逻辑回归

⑤ 高斯判别分析

⑥ 朴素贝叶斯分类器

背景(Background)

- 频率派 → 统计机器学习　核心是线性回归.
- 贝叶斯派 → 概率图模型

Linear Regression

① 线性
- 属性线性
- 系数线性
- 结果线性 → 直接将线性组合作为结果

$f(w,b)=w^Tx+b$. 关于x是线性的，关于w^T和b也是线性的。（b → 可以合并到W中）

② 全局性　对数据的拟合是全局的，不会局部拟合

③ 数据未加工　数据没有经过降维等处理.

将线性回归这三个性质分别破坏后，就形成了统计机器学习的大框架.

① 线性 ⇒ 非线性
- 属性非线性　特征转换 → 多项式回归 $w_1x_1^2+w_2x_2^2+\cdots$
- 系数非线性　神经网络. 这里的系数非线性指的是系数会变化
  - 系数其实不存在非线性，w^2也只是系数，可以令$\beta=w^2$保持线性
  - 例：神经网络初始化参数不同，最终得到的系数也不同.
- 结果非线性　线性分类 → $f(w^Tx+b)$.
  $f(\cdot)$是非线性函数，使得结果输出为非线性.

② 全局性 ⟹ 局部性

线性样条回归，将数据分段拟合，具有局部性.

决策树，将样本空间划分为一个个子空间，使其局部具有不同决策.

③ 数据未加工 ⟹ 处理数据. PCA. 流形

线性回归 $\xrightarrow[\text{降维}]{\text{激活函数}}$ 线性分类

$y=f(w^Tx+b)$，$y\in\{0,1\}$ 或 $[0,1]$
$\{0,1\}$ 硬分类　$[0,1]$ 软分类

$f(\cdot)\to$ activation function

$\hookrightarrow w^Tx+b \longmapsto \{0,1\}$ 或 $[0,1]$

$f^{-1}(\cdot)\to$ link function

$\hookrightarrow \{0,1\}$ 或 $[0,1] \longmapsto w^Tx+b$.

线性分类
- 硬分类 $y\in\{0,1\}$
  - 线性判别分析 → fisher 判别分析
  - 感知机
- 软分类 $y\in[0,1]$
  - 判别式：Logistic Regression
    直接建模 $P(Y|X)$
  - 生成式：
    - Naive Bayes ↗ 离散
    - Gaussian Discriminant Analysis ↗ 连续

$P(Y|X)=\frac{P(X|Y)P(Y)}{P(X)}\propto P(X|Y)P(Y)\xrightarrow{MAP}P(Y=0|X),P(Y=1|X)$

感知机 (Perceptron)

思想：$\boxed{\text{错误驱动}}$

模型：$f(x) = \text{sign}(w^T x)$，$x \in \mathbb{R}^p$. $w \in \mathbb{R}^p$，$\text{sign}(a) = \begin{cases} +1, & a \geq 0 \\ -1, & a < 0 \end{cases}$

策略：Loss Function 被错误分类的样本的数量.

$L'(w) = \sum_{i=1}^{N} I\{\underline{y_i w^T x_i < 0}\}$.

错误分类时，y_i 与 $w_i x_i$ 异号.

但是 $w \to w + \Delta w$，I 就会从0阶跃到1或从1阶跃到0

即 $L'(w)$ 不可导.

∴不用 $L'(w)$ 作为损失函数.

令 $D = \{\text{误分类的样本}\}$

则 $L(w) = \sum_{x_i \in D} -y_i w^T x_i$，将其作为损失函数.

$\nabla_w L(w) = \sum_{x_i \in D} -y_i x_i$

参数的更新采用梯度下降法：全体样本

$w^{(t+1)} \leftarrow w^{(t)} - \lambda \left(\sum_{x_i \in D} -y_i x_i \right) = w^{(t)} + \lambda \sum_{x_i \in D} y_i x_i$

或随机梯度下降：单个样本.

$w^{(t+1)} \leftarrow w^{(t)} - \lambda(-y_i x_i) = w^{(t)} + \lambda y_i x_i$

线性可分的数据感知机算法才能收敛.

若线性不可分，则需要使用感知机的变形 pocket algorithm. 允许有错误率

线性判别分析 (Linear Discriminant Analysis)

设 $X = (x_1, x_2, \cdots, x_N)^T \in \mathbb{R}^{N\times p}$

$Y = (y_1, y_2, \cdots, y_N)^T \in \mathbb{R}^{N\times 1}$, $y_i \in \{\overset{C_1}{\overset{\uparrow}{+1}}, \overset{C_2}{\overset{\uparrow}{-1}}\}$

$X_{C_1} = \{x_i \mid y_i = +1\}$, X_{C_1} 集合大小为 N_1

$X_{C_2} = \{x_i \mid y_i = -1\}$. X_{C_2} 集合大小为 N_2

$N_1 + N_2 = N$.

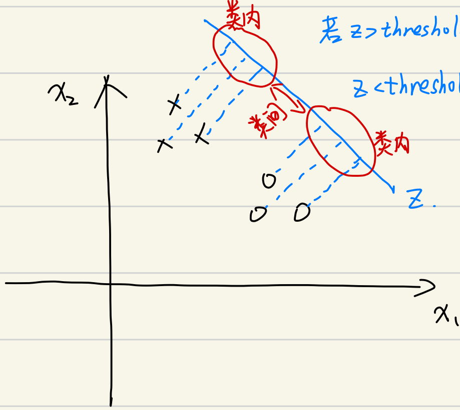
思想：类内小，类间大，运用降维的方式.

$\boxed{\text{找一条合适的 } z \text{ 轴}}$

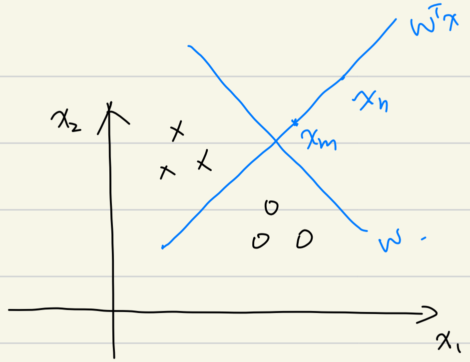
将二维降到1维

若 $z >$ threshold, $+1$

$z <$ threshold, -1



假设超平面 $w^Tx=0$ 将两类样本分割开，
求证明，投影的轴即为超平面的法向量 w.



证明：由于 $w^Tx=0$.

任取超平面上两点

得 $w^Tx_m=0$，$w^Tx_n=0$

则 $w^T(x_m-x_n)=0$.

由于 w 与 (x_m-x_n) 内积为0. 则 w 垂直于 (x_m-x_n)

由于 x_m, x_n 是任意选取的，则 w 为超平面 $w^Tx=0$ 的法向量

$\boxed{\text{规定超平面 } w^Tx=0 .\quad \|w\|=1.}$

则样本点 x_i 在超平面法向量上的投影长度为 $z_i=\underline{w^Tx_i}$ 内积

则均值 $\bar{z}=\frac{1}{N}\sum_{i=1}^{N}z_i=\frac{1}{N}\sum_{i=1}^{N}w^Tx_i$

方差 $S=\frac{1}{N}\sum_{i=1}^{N}(z_i-\bar{z})(z_i-\bar{z})^T$

对于 C_1 类投影，均值 $\bar{z}_1=\frac{1}{N_1}\sum_{x_i\in X_{c_1}}w^Tx_i$

方差 $S_{z_1}=\frac{1}{N_1}\sum_{x_i\in X_{c_1}}(w^Tx_i-\bar{z}_1)(w^Tx_i-\bar{z}_1)^T$

对于 C_2 类投影，均值 $\bar{z}_2=\frac{1}{N_2}\sum_{x_i\in X_{c_2}}w^Tx_i$

方差 $S_{z_2}=\frac{1}{N_2}\sum_{x_i\in X_{c_2}}(w^Tx_i-\bar{z}_2)(w^Tx_i-\bar{z}_2)^T$

这里需要注意的是，z_i 为投影后的向量长度.

同时，对于C_1类样本 $\bar{x}_{c_1}=\frac{1}{N_1}\sum_{x_i\in x_{c_1}} x_i$

方差：$S_{c_1}=\frac{1}{N_1}\sum_{x_i\in x_{c_1}}(x_i-\bar{x}_{c_1})(x_i-\bar{x}_{c_1})^T$

对于C_2类样本：$\bar{x}_{c_2}=\frac{1}{N_2}\sum_{x_i\in x_{c_2}} x_i$

方差：$S_{c_2}=\frac{1}{N_2}\sum_{x_i\in x_{c_2}}(x_i-\bar{x}_{c_2})(x_i-\bar{x}_{c_2})^T$

则 类间方差：$S_b=(\bar{x}_{c_1}-\bar{x}_{c_2})(\bar{x}_{c_1}-\bar{x}_{c_2})^T$ between - class

类内方差：$S_w=S_{c_1}+S_{c_2}$. within - class

注意，这里的 $\bar{x}_{c_1}$，S_{c_1}，$\bar{x}_{c_2}$，S_{c_2}，S_b，S_w 都不是一维的，是针对原始空间来说的，

为了使投影后实现"类内小，类间大"，设目标函数：

$$J(w)=\frac{(\bar{z}_1-\bar{z}_2)^2}{S_{z_1}+S_{z_2}}$$ 则 $\hat{w}=\arg\max_w J(w)$

对于分子：
$$
\begin{aligned}
(\bar{z}_1-\bar{z}_2)^2 &= (\bar{z}_1-\bar{z}_2)(\bar{z}_1-\bar{z}_2)^T \\
&= \left(\frac{1}{N_1}\sum_{x_i\in x_{c_1}} w^T x_i-\frac{1}{N_2}\sum_{x_i\in x_{c_2}} w^T x_i\right)\left(\frac{1}{N_1}\sum_{x_i\in x_{c_1}} w^T x_i-\frac{1}{N_2}\sum_{x_i\in x_{c_2}} w^T x_i\right)^T \\
&= w^T\left(\frac{1}{N_1}\sum_{x_i\in x_{c_1}} x_i-\frac{1}{N_2}\sum_{x_i\in x_{c_2}} x_i\right)\left(\frac{1}{N_1}\sum_{x_i\in x_{c_1}} x_i-\frac{1}{N_2}\sum_{x_i\in x_{c_2}} x_i\right)^T w \\
&= w^T(\bar{x}_{c_1}-\bar{x}_{c_2})(\bar{x}_{c_1}-\bar{x}_{c_2})^T w \\
&= w^T S_b w.
\end{aligned}
$$

对于分母：$S_{z_1}+S_{z_2}=\frac{1}{N_1}\sum_{x_i\in X_{c_1}}\left(w^Tx_i-\frac{1}{N_1}\sum_{x_i\in X_{c_1}}w^Tx_i\right)\left(w^Tx_i-\frac{1}{N_1}\sum_{x_i\in X_{c_1}}w^Tx_i\right)^T$

$$+\frac{1}{N_2}\sum_{x_i\in X_{c_2}}\left(w^Tx_i-\frac{1}{N_2}\sum_{x_i\in X_{c_2}}w^Tx_i\right)\left(w^Tx_i-\frac{1}{N_2}\sum_{x_i\in X_{c_2}}w^Tx_i\right)^T$$

$$=\frac{1}{N_1}\sum_{x_i\in X_{c_1}}w^T(x_i-\bar{x}_{c_1})(x_i-\bar{x}_{c_1})^Tw$$

$$+\frac{1}{N_2}\sum_{x_i\in X_{c_2}}w^T(x_i-\bar{x}_{c_2})(x_i-\bar{x}_{c_2})^Tw$$

$$=w^T\left[\frac{1}{N_1}\sum_{x_i\in X_{c_1}}(x_i-\bar{x}_{c_1})(x_i-\bar{x}_{c_1})^T\right]w$$

$$+w^T\left[\frac{1}{N_2}\sum_{x_i\in X_{c_2}}(x_i-\bar{x}_{c_2})(x_i-\bar{x}_{c_2})^T\right]w$$

$$=w^TS_{c_1}w+w^TS_{c_2}w$$

$$=w^T(S_{c_1}+S_{c_2})w$$

$$=w^TS_ww$$

则目标函数：$J(w)=\dfrac{w^TS_bw}{w^TS_ww}$

$$\frac{\partial J(w)}{\partial w}=2S_bw\,(w^TS_ww)^{-1}+(-1)\,w^TS_bw\,(w^TS_ww)^{-2}\cdot 2S_ww$$

$$=0$$

则 $2S_b w\left(\boxed{w^T S_w w}\right)^{-1} = w^T S_b w \left(\boxed{w^T S_w w}\right)^{-2} \cdot 2 S_w w$ 都为实数.

$$S_b w (w^T S_w w) = w^T S_b w S_w w \Leftarrow \text{两边同乘} (w^T S_w w)^2$$

$$\frac{w^T S_w w}{w^T S_b w} S_b w = S_w w$$

$$w = \boxed{\frac{w^T S_w w}{w^T S_b w}} S_w^{-1} S_b w$$

一维.

由于 $w^T S_w w$ 和 $w^T S_b w$ 都是一维实数，则

由于 w 只需考虑其方向，大小不用考虑（前面约束 $\|w\|=1$，就算前面没有约束，最后可以 scaling 成 1）.

所以可以直接舍去一维实数.

$$w \propto S_w^{-1} S_b w.$$

$$\propto S_w^{-1}(\bar{x}_{c_1}-\bar{x}_{c_2})\underbrace{(\bar{x}_{c_1}-\bar{x}_{c_2})^T w}_{\text{一维}}$$

$$\boxed{\propto S_w^{-1}(\bar{x}_{c_1}-\bar{x}_{c_2})}$$

至此，求得 w 的方向. 进一步，如果 C_1 类与 C_2 类各向同性（协方差矩阵为对角矩阵，且对角线元素相同），

那么 $S_w = \lambda I$. 即 $w \propto (\bar{x}_{c_1}-\bar{x}_{c_2})$

综上，求得投影的方向，以达到"类内小，类间大"的要求.

逻辑回归（Logistic Regression）

线性回归的结果是线性组合：$w^T x$

逻辑回归的结果是类别：$\{0,1\}$.

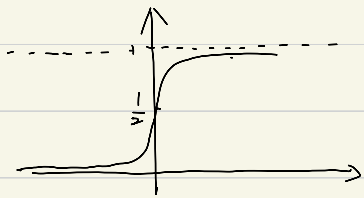
用激活函数将 $w^T x$ 映射到 $\{0,1\}$.

激活函数：sigmoid $\to$ $\sigma(z) = \dfrac{1}{1+e^{-z}}$

$\lim\limits_{z\to\infty} \sigma(z) = 1$

$\lim\limits_{z\to-\infty} \sigma(z) = 0$

$\lim\limits_{z\to 0} \sigma(z) = \frac{1}{2}$.



则 $\sigma: \mathbb{R} \longmapsto (0,1)$

$w^T x \longmapsto p$.

令 $\begin{cases} p_1 = P(y=1|x) = \sigma(w^T x) = \dfrac{1}{1+e^{-w^T x}}, & y=1 \\ p_0 = P(y=0|x) = 1 - P(y=1|x) = \dfrac{e^{-w^T x}}{1+e^{-w^T x}}, & y=0 \end{cases}$

则 $P(y|x) = p_1^y p_0^{1-y}$

用极大似然估计就可以直接估计 w.

$$
\begin{aligned}
\text{MLE: } \hat{w} &= \arg\max_{w} \log P(Y|X) \\
&= \arg\max_{w} \log \prod_{i=1}^{N} p(y_i|x_i) \\
&= \arg\max_{w} \sum_{i=1}^{N} \log p(y_i|x_i) \\
&= \arg\max_{w} \sum_{i=1}^{N} \left[y_i \log p_1 + (1-y_i) \log p_0 \right] \\
&= \arg\max_{w} \underbrace{\sum_{i=1}^{N} \left[y_i \log p(y=1|x_i) + (1-y_i) \left(\log(1-p(y=1|x_i)\right) \right]}_{(-1)\times \text{cross entropy}}.
\end{aligned}
$$

max MLE $\Rightarrow$ min Loss Function

$$
= \arg\max_{w} \sum_{i=1}^{N} \left[-y_i \log(1+e^{-w^T x_i}) - (1-y_i)\left(w^T x_i + \log(1+e^{-w^T x_i}) \right) \right]
$$

$$
= \arg\min_{w} \underbrace{\sum_{i=1}^{N} \left[y_i \left(\log(1+e^{w^T x_i}) - w^T x_i \right) + (1-y_i) \log(1+e^{w^T x_i}) \right]}_{\mathcal{L}(w)}.
$$

$$
\begin{aligned}
\frac{\partial \mathcal{L}(w)}{\partial w} &= \sum_{i=1}^{N} \left[y_i \left(\frac{x_i e^{w^T x_i}}{1+e^{w^T x_i}} - x_i \right) + (1-y_i) \left(\frac{x_i e^{w^T x_i}}{1+e^{w^T x_i}} \right) \right] \\
&= \sum_{i=1}^{N} \left[y_i \frac{x_i e^{w^T x_i}}{1+e^{w^T x_i}} - y_i x_i + \frac{x_i e^{w^T x_i}}{1+e^{w^T x_i}} - y_i \frac{x_i e^{w^T x_i}}{1+e^{w^T x_i}} \right] \\
&= \sum_{i=1}^{N} \left(\frac{x_i e^{w^T x_i}}{1+e^{w^T x_i}} - y_i x_i \right)
\end{aligned}
$$

则使用梯度下降法，即可求得 w 的解：

$$
w^{(t+1)} \leftarrow w^{(t)} - \underbrace{\lambda}_{\text{步长}} \frac{\partial \mathcal{L}(w)}{\partial w}.
$$

使用牛顿法也可以，只不过需要求二阶导数。

高斯判别分析

(Gaussian Discriminant Analysis)

Data: $\{(x_i, y_i)\}$ $x_i \in \mathbb{R}^p$, $y_i \in \{0,1\}$

高斯判别分析属于前面提到的生成式模型.

对于判别式模型, 是求 $p(y|x)$ 的大小. 对 $p(y|x)$ 建模

对于生成式模型, 是求 $p(y=0|x)$ 和 $p(y=1|x)$ 哪个大 对 $p(x,y)$ 建模

因此, 根据贝叶斯公式: $p(y|x) = \dfrac{p(x|y)\,p(y)}{p(x)}$ 边缘分布与 y 无关, 故可舍去.

可得 $p(y|x) \propto \underline{p(x|y)\,p(y)}$ 只需要考虑这个联合分布即可. $p(x,y)$

MAP: $\hat{y} = \underset{y\in\{0,1\}}{\arg\max}\, p(y|x) = \underset{y}{\arg\max}\, p(y)\cdot p(x|y)$

由于 y 只有两个取值, 故此时可以看作是伯努利分布

$y \sim \text{Bernoulli}(\phi)$

y	0	1
p	$1-\phi$	ϕ

$\Rightarrow p(y) = \phi^y (1-\phi)^{1-y}$

$x|y=1 \sim N(\mu_1, \Sigma)$

$x|y=0 \sim N(\mu_2, \Sigma)$

高斯判别分析的假设

$\Rightarrow x|y \sim \underbrace{N(\mu_1, \Sigma)}_{P_1}{}^{y} \cdot \underbrace{N(\mu_2, \Sigma)}_{P_2}{}^{1-y} = P_1^{\,y} P_2^{\,1-y}$

联合概率之似然函数：$L(\theta) = \log \prod_{i=1}^{N} p(x_i, y_i)$

$$= \sum_{i=1}^{N} \log p(x_i|y_i) p(y_i)$$

$$= \sum_{i=1}^{N} [\log p(x_i|y_i) + \log p(y_i)]$$

$$= \sum_{i=1}^{N} [\log P_1^{y_i} P_2^{1-y_i} + \log \phi^{y_i}(1-\phi)^{1-y_i}]$$

$$= \sum_{i=1}^{N} [y_i \log P_1 + (1-y_i)\log P_2 + \log \phi^{y_i}(1-\phi)^{1-y_i}]$$

其中，$\theta = (\phi, \mu_1, \mu_2, \Sigma)$

则 MLE：$\theta = \underset{\theta}{\operatorname{argmax}}\, L(\theta)$

对于 ϕ：
$$\frac{\partial L(\theta)}{\partial \phi} = \frac{\partial}{\partial \phi}[\sum_{i=1}^{N} \log \phi^{y_i}(1-\phi)^{1-y_i}]$$

$$= \sum_{i=1}^{N} [y_i \log \phi + (1-y_i) \log(1-\phi)]$$

$$= \sum_{i=1}^{N} [\frac{y_i}{\phi} - \frac{1-y_i}{1-\phi}]$$

令 $\frac{\partial L(\theta)}{\partial \phi} = 0$.

则 $\sum_{i=1}^{N} [y_i - y_i\phi - \phi + y_i\phi] = 0$

$$\phi = \frac{1}{N}\underbrace{\sum_{i=1}^{N} y_i}_{N_1} = \frac{N_1}{N}$$

对于 μ_1. $\frac{\partial L(\theta)}{\partial \mu_1} = \frac{\partial}{\partial \mu_1}\left[\sum_{i=1}^{N} y_i \log p_1\right]$

其中 $p_1 = \frac{1}{(2\pi)^{\frac{1}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(x-\mu_1)^T \Sigma^{-1}(x-\mu_1)\right\}$.

则. $\frac{\partial L(\theta)}{\partial \mu_1} = \frac{\partial}{\partial \mu_1} \sum_{i=1}^{N} -\frac{1}{2} y_i (x_i-\mu_1)^T \Sigma^{-1}(x_i-\mu_1)$

$$= \frac{\partial}{\partial \mu_1} \sum_{i=1}^{N} -\frac{1}{2} y_i (x_i^T\Sigma^{-1} - \mu_1^T\Sigma^{-1})(x_i-\mu_1)$$

$$= \frac{\partial}{\partial \mu_1} \sum_{i=1}^{N} -\frac{1}{2} y_i \left[x_i^T\Sigma^{-1}x_i - 2x_i^T\Sigma^{-1}\mu_1 + \mu_1^T\Sigma^{-1}\mu_1\right].$$

$$= \sum_{i=1}^{N} -\frac{1}{2} y_i (-2\Sigma^{-1}x_i + 2\Sigma^{-1}\mu_1)$$

令 $\frac{\partial L(\theta)}{\partial \mu_1} = 0$

则 $\sum_{i=1}^{N} y_i (\Sigma^{-1}x_i - \Sigma^{-1}\mu_1) = 0$

$$\mu_1 \sum_{i=1}^{N} y_i = \sum_{i=1}^{N} y_i x_i$$

$$\mu_1 = \frac{\sum_{i=1}^{N} y_i x_i}{\underbrace{\sum_{i=1}^{N} y_i}_{N_1}}$$

同理: 对于 μ_2: $\mu_2 = \frac{\sum_{i=1}^{N}(1-y_i)x_i}{\underbrace{\sum_{i=1}^{N}(1-y_i)}_{N_2}}$.

$\rightarrow N_1 + N_2 = N$.

对于Σ：

$$\frac{\partial L(\theta)}{\partial \Sigma} = \frac{\partial}{\partial \Sigma}\Big[\sum_{i=1}^{N} y_i \log \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} \exp\Big\{-\frac{1}{2}(x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1)\Big\}$$

$$+\sum_{i=1}^{N}(1-y_i)\log\frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}}\exp\Big\{-\frac{1}{2}(x_i-\mu_2)^T\Sigma^{-1}(x_i-\mu_2)\Big\}\Big]$$

$$= \frac{\partial}{\partial \Sigma}\Big[\sum_{i=1}^{N} y_i\Big(-\frac{1}{2}\log|\Sigma| - \frac{1}{2}(x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1)\Big)$$

$$+\sum_{i=1}^{N}(1-y_i)\Big(-\frac{1}{2}\log|\Sigma| - \frac{1}{2}(x_i-\mu_2)^T\Sigma^{-1}(x_i-\mu_2)\Big)\Big]$$

$$= \frac{\partial}{\partial \Sigma}\Big[\sum_{i=1}^{N} -\frac{1}{2}\log|\Sigma| + \sum_{i=1}^{N} -\frac{1}{2}y_i(x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1)$$

$$+\sum_{i=1}^{N} -\frac{1}{2}(1-y_i)(x_i-\mu_2)^T\Sigma^{-1}(x_i-\mu_2)\Big].$$

令 $C_1 = \{x_i \mid y_i = 1, i = 1, 2, \cdots, N\}$

$C_2 = \{x_i \mid y_i = 0, i = 1, 2, \cdots, N\}$.

且C_1中元素的个数为N_1，C_2中为N_2. $N_1 + N_2 = N$.

则

$$\frac{\partial L(\theta)}{\partial \Sigma} = \frac{\partial}{\partial \Sigma}\Big[-\frac{1}{2}N\log|\Sigma| - \frac{1}{2}\sum_{x_i\in C_1}(x_i-\mu_1)^T\Sigma^{-1}(x_1-\mu_1)$$

$$-\frac{1}{2}\sum_{x_i\in C_2}(x_i-\mu_2)^T\Sigma^{-1}(x_i-\mu_2)\Big]$$

$$= -\frac{1}{2}N\cdot\Sigma^{-1} + \frac{1}{2}\sum_{x_i\in C_1}(x_i-\mu_1)(x_i-\mu_1)^T\cdot\Sigma^{-2}$$

$$+\frac{1}{2}\sum_{x_i\in C_2}(x_i-\mu_2)(x_i-\mu_2)^T\Sigma^{-2}$$

$$\frac{\partial|A|}{\partial A} = |A|A^{-1}$$

$$\frac{\partial x^TAx}{\partial A} = xx^T$$

A对称

$$\frac{\partial A^{-1}}{\partial A} = -A^{-2}$$

其中 $S_1 = \frac{1}{N_1}\sum_{x_i\in C_1}(x_i-\mu_1)(x_i-\mu_1)^T$ C_1的样本方差

$S_2 = \frac{1}{N_2}\sum_{x_i\in C_2}(x_i-\mu_2)(x_i-\mu_2)^T$ C_2的样本方差

则 $\frac{\partial L(\theta)}{\partial \Sigma} = -\frac{1}{2}N\cdot\Sigma^{-1}+\frac{1}{2}N_1S_1\cdot\Sigma^{-2}+\frac{1}{2}N_2S_2\cdot\Sigma^{-2}=0$.

则 $N\Sigma = N_1S_1+N_2S_2$

$\Sigma = \frac{1}{N}(N_1S_1+N_2S_2)$.

原白板系列中用到的方法和我写的这个稍微有点不同.

原视频中. 将 $(x-\mu)^T\Sigma^{-1}(x-\mu)$ 看作是 $tr((x-\mu)^T\Sigma^{-1}(x-\mu))$

一个数的迹是它本身.

并利用

$tr(AB)=tr(BA)$

$tr(ABC)=tr(CAB)=tr(BCA)$

$\frac{\partial tr(AB)}{\partial A} = B^T$

三个公式求它的导数

其实实质上还是一样的

只不过我自己手推了 x^TAx 对 A 的导数, 直接使其一步到位.

原视频中的感觉更巧妙.

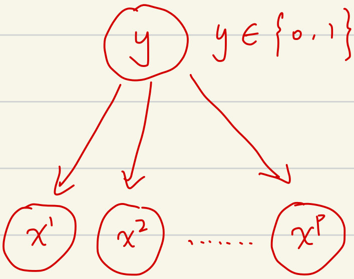
朴素贝叶斯（Naive Bayes）

思想：朴素贝叶斯假设（条件独立性假设）就是为了简化运算

最简单的概率图，也是最简单的概率生成模型

（有向图）.

相同类别中，不同特征之间是独立的.



$y \in \{0,1\}$

$x \in \mathbb{R}^p$

$x^i \perp x^j \mid y \quad (i \neq j)$

即 $p(x|y) = \prod_{i=1}^{p} p(x^i|y)$

则对于给定数据：

Data: $\{(x_i, y_i)\}$, $x_i \in \mathbb{R}^p$, $y_i \in \{0,1\}$.

$$\text{MAP}: \hat{y} = \arg\max_y p(y|x)$$

$$= \arg\max_y \frac{p(x|y)\,p(y)}{p(x)}$$

$$= \arg\max_y p(x|y)\,p(y)$$

对于 $p(y)$：2分类：$y \sim$ Bernoulli Distribution

多分类：$y \sim$ Categorical Distribution

对于 $p(x|y)$：$p(x|y) = \prod_{i=1}^{p} p(x^i|y)$.

当 x 离散：$x^i \sim$ Categorical Distribution

当 x 连续：$x^i \sim N(\mu_i, \sigma_i^2)$

对上述提到的分布进行简单介绍：

Categorical Distribution：

这是将伯努利分布推广到多分类的结果：意味着 p 个 x_i 中，只有一个为1 其他为0

设 X 有 p 个取值：$(x_1, x_2, \cdots, x_p)$，$\boxed{x_i \in \{0,1\} \text{ 且 } \sum x_i = 1.}$

每个取值对应的概率：$(\alpha_1, \alpha_2, \cdots, \alpha_p)$. $\alpha_i \in [0,1]$, $\sum \alpha_i = 1$

则 $P(X = x_i) = \prod_{i=1}^{p} \alpha_i^{x_i}$

Bernoulli Distribution $\xrightarrow{n次}$ Binormial Distribution

$\downarrow$ 多分类.

Categorical Distribution $\xrightarrow{n次}$ Multinormial Distribution.

下面将假设 $y \sim$ Bernoulli Distribution

$x^i \sim$ Categorical Distribution

对参数进行估计.

由于分布指定，参数为确定的数，则采用MLE.

$$\hat{y} = \arg\max_{y} p(x, y).$$

$$
\begin{aligned}
L(\beta) &= \log \prod_{i=1}^{N} P(x_i, y_i) \\
&= \sum_{i=1}^{N} \log P(x_i|y_i)\, p(y_i) \\
&= \sum_{i=1}^{N} \Big[\log p(y_i) + \log \prod_{j=1}^{p} p(x_i^j|y_i)\Big] \\
&= \sum_{i=1}^{N} \Big[\log p(y_i) + \sum_{j=1}^{p} \log p(x_i^j|y_i)\Big].
\end{aligned}
$$

由于 $y \sim$ Bernoulli Distribution $\Rightarrow p(y) = \phi^{y}(1-\phi)^{1-y}$

$x^j \sim$ Categorical Distribution $\Rightarrow p(x^j|y) = \prod_{k=1}^{m} \theta_k^{x^{j,k}}$ $\quad \sum_{k=1}^{m} \theta_k = 1$

意思是x第j个特征取第k个值，$x^{j,k} \in \{0,1\}$，0代表取，1代表不取.

$$
\begin{aligned}
\text{则 } L(\beta) &= \sum_{i=1}^{N} \Big[\log \phi^{y}(1-\phi)^{1-y} + \sum_{j=1}^{p} \log \prod_{k=1}^{m} \theta_k^{x_i^{j,k}}\Big] \\
&= \sum_{i=1}^{N} \Big[\log \phi^{y_i}(1-\phi)^{1-y_i} + \sum_{j=1}^{p} \sum_{k=1}^{m} x_i^{j,k} \log \theta_k\Big]
\end{aligned}
$$

令 $\beta = (\phi, \theta_k, \cdots)$

注意. 每个特征都有K个θ_K，且不同特征的θ_K都不同. 但这里由于公式中的θ_K上下标太多了，且每个特征的θ_K求解方式都是一样的. 故没有给他区分

$$\beta = \arg\max_{\beta} L(\beta).$$

对于 ϕ.

$$\frac{\partial L(\beta)}{\partial \phi} = \frac{\partial}{\partial \phi} \sum_{i=1}^{N} \log \phi^{y_i}(1-\phi)^{1-y_i} = 0$$

$$\Rightarrow \boxed{\phi = \frac{1}{N}\sum_{i=1}^{N} y_i}$$

对于 θ_k.

$$L(\beta) = \sum_{i=1}^{N}\sum_{j=1}^{P}\sum_{k=1}^{m} x_i^{j,m} \log \theta_k .$$

由于每个特征的 θ_k 都是不同的.

其实这一步要到后面求偏导才可以把 $\sum_{j=1}^{P}$ 消去，但是提前做也影响不了什么.
因为本质上来说，我当前只对该特征求解.
不关其他特征什么事.

则 $L(\beta) = \sum_{i=1}^{N}\sum_{k=1}^{m} x_i^m \log \theta_k .$

由于 $x_i^m \in \{0,1\}$. 设对于整个 Data，在该特征中取第 k 个值的样本数为 N_k.

则 $L(\beta) = \sum_{k=1}^{m} N_k \log \theta_k$ 且 $\sum_{k=1}^{m} \theta_k = 1$.

用拉格朗日乘子法. 入 $F = \sum_{k=1}^{m} N_k \log \theta_k + \lambda(1-\sum_{k=1}^{m}\theta_k)$.

$$\begin{cases} \dfrac{\partial F}{\partial \theta_k} = \dfrac{N_k}{\theta_k} - \lambda = 0 & ① \\[2ex] \dfrac{\partial F}{\partial \lambda} = 1 - \sum_{k=1}^{m} \theta_k = 0 & ② \end{cases}$$

由①，得 $\theta_k = \frac{N_k}{\lambda}$ ③

将③代入②，得：$\lambda = N$.

则 $\boxed{\theta_k = \frac{N_k}{N}}$. 对其他特征求法相同.