1. TF-IDF 词频-逆文件频率

> 字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。
>
> 总结就是, 一个词语在一篇文章中出现次数越多, 同时在所有文档中出现次数越少, 越能够代表该文章.

词频 $TF = \frac{词w在文档d中出现次数}{文档d总词数}$

逆文档频率 $IDF = log(\frac{文档总数+1}{包含词w的文档数+1}) + 1$

注：分母+1是为了防止分母为0；分子+1是为了防止分母大于分子导致结果为负数；尾部+1是为了防止结分子=分母，导致IDF为0，影响最终结果

2. 线性回归 Linear Regression

预测函数

$f(x) = \theta^T x + b$, 另θ包含b, 则公式简化为 $f(x) = \theta^T x$

误差：$\epsilon_i = f(x_i) - y_i$

前提：各个样本点是独立的，ϵ是独立同分布，根据独立同分布中心极限定理，ϵ 服从均值$\mu = 0$，方差为σ^2的高斯分布（正态分布$f(x) = \frac{1}{\sqrt{2\pi}\sigma}exp(-\frac{(x-\mu)^2)}{2\sigma^2}))$

所以：$p(\epsilon_i) = \frac{1}{\sqrt{2\pi}\sigma}exp(-\frac{\epsilon_i^2}{2\sigma^2})$, 可以看作是给定了$\theta$和 x_i情况下的y_i取值

即：$p(y_i|x_i;\theta) = \frac{1}{\sqrt{2\pi}\sigma}exp(-\frac{(\theta^T x_i - y_i)^2}{2\sigma^2})$

极大似然估计

$L(\theta) = argmax \prod p(y_i|x_i;\theta)$

$= argmax \sum logP(y_i|x_i;\theta)$ (log化)

$= argmax \sum log(\frac{1}{\sqrt{2\pi}\sigma}exp(-\frac{(\theta^T x_i - y_i)^2}{2\sigma^2}))$

$= argmax \sum [log(\frac{1}{\sqrt{2\pi}\sigma}) + logexp(-\frac{(\theta^T x_i - y_i)^2}{2\sigma^2})]$

$= argmax \sum [log(\frac{1}{\sqrt{2\pi}\sigma}) - \frac{1}{2\sigma^2}(\theta^T x_i - y_i)^2]$

所以问题转化为:

$argmin \sum (\theta^T x_i - y_i)^2$

平方损失函数

$J(\theta) = \frac{1}{2}\sum(\theta^T x_i - y_i)^2$

3. 感知机（线性分类）

预测函数

$f(x) = sign(\theta^T x + b)$, 另$\theta$包含$b$, 则公式简化为 $f(x) = sign(\theta^T x)$

0-1损失函数

4. 逻辑斯蒂回归（线性分类）

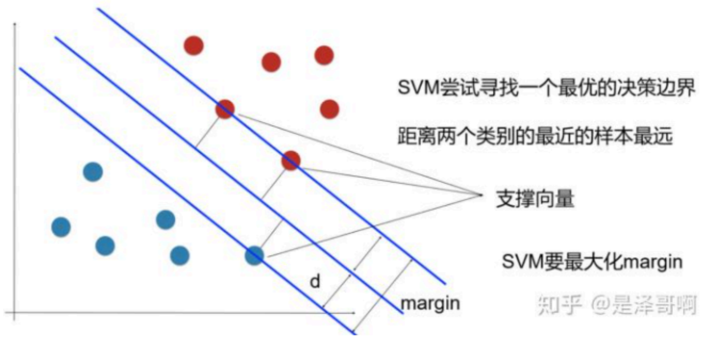
预测函数

$f(x) = sigmoid(\theta^T x) = \frac{1}{1+e^{-\theta^T x}}$

对数损失函数（二分类交叉熵损失函数）

$L(Y, f(X)) = -logP(Y|X)$

5. SVM



目标函数

分隔线：

$y = ax + b => ax - y + b = 0 => [a, -1]^T[x, y] + b = 0$

令系数 $w = [a, -1], x = [x, y]$

$=> w^T x + b = 0$

点到分隔线的距离：

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

假设蓝色点是1，红色点是-1，所以，我们的目标是：

$\frac{w^Tx+b}{||w||} >= d(y == 1)$

$\frac{w^Tx+b}{||w||} <= -d(y == -1)$

可以转化为:

$y\frac{w^Tx+b}{||w||} >= d$

又因为$||w||$和d都为常量

$y\frac{w^Tx+b}{d||w||} >= 1$

假设$d||w|| == 1$(对推导无影响)

$y(w^Tx+b) >= 1$

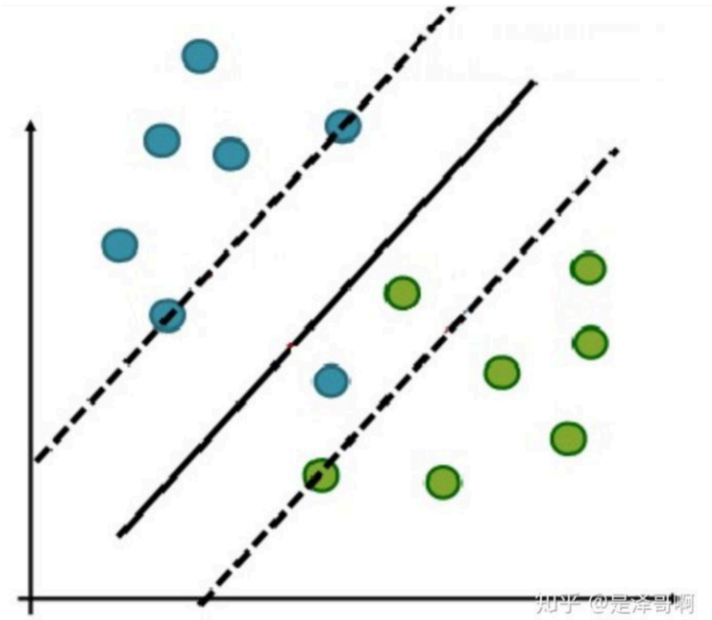
损失函数

我们需要让d越大越好，即让$||w||$越小越好，所以：

$min\frac{1}{2}||w||$，且$(y(w^Tx+b) >= 1)$

用拉格朗日乘数发求解等是约束优化问题

软间隔



允许部分点：$w^Tx+b < 1$

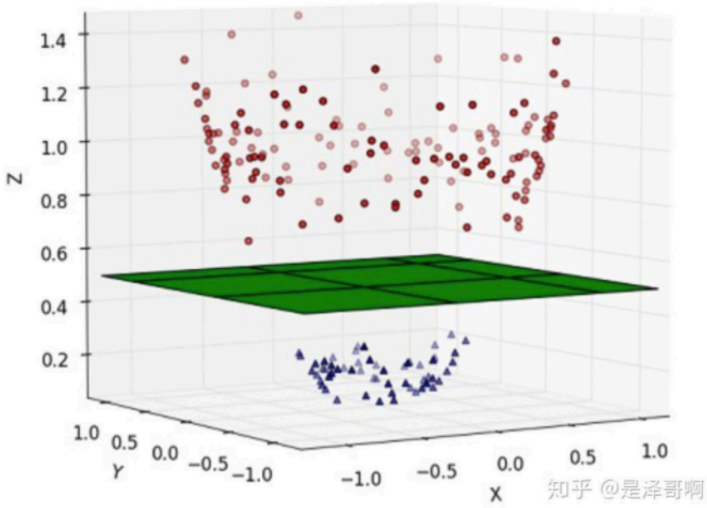
引入松弛变量$\xi_i >= 0$

目标函数变为：$y(w^Tx+b)+\xi >= 1$

损失函数变为：$min\frac{1}{2}||w||+\sum\xi$，且$(y(w^Tx+b)-\xi >= 1)$

线性不可分

将样本映射到高纬空间，用超平面分割



使用核函数能够减少高纬映射计算量

常见的核函数有：

线性核函数、多项式核函数、高斯核函数

优缺点

- 可解释行强
- 只适合小批量任务，效率低

6. 朴素贝叶斯

> https://zhuanlan.zhihu.com/p/26262151

求解样本X属于哪个类别，即求解出，X输出各个类别的概率$P(Y_i|X)$，看哪个最大。

比如， 评价一个西瓜X有以下几个独立指标$[x^1\text{：色泽，} x^2\text{：声音，} x^3\text{：大小}]$，$Y \in \{0:\text{坏瓜}, 1:\text{好瓜}\}$。

0-1 损失函数

分类正确，损失为0，分类错误，损失为1

$$L(Y, f(X)) = \begin{cases} 1, Y \neq f(X) \\ 0, Y = f(X) \end{cases}$$

所以：

$$Loss = argminL(y, f(X))P(Y = y|X = X_i)$$

$$= argminP(Y \neq y|X = X_i)$$

$= argmin\ (1 - P(Y = y|X = X_i))$

$= argmax P(Y = y|X = X_i)$

所以，最终目标是让后验概率 $P(Y = y|X = X_i)$最大，则整体损失最小

$P(Y = y|X = X_i) = \frac{P(X=X_i,Y=y)}{P(X=X_i)}$ （分母可用全概率公式转化，并且分母可以忽略）

则，等价于目标是找到一个合适的y让联合概率分布 $P(X = X_i, Y = y)$最大

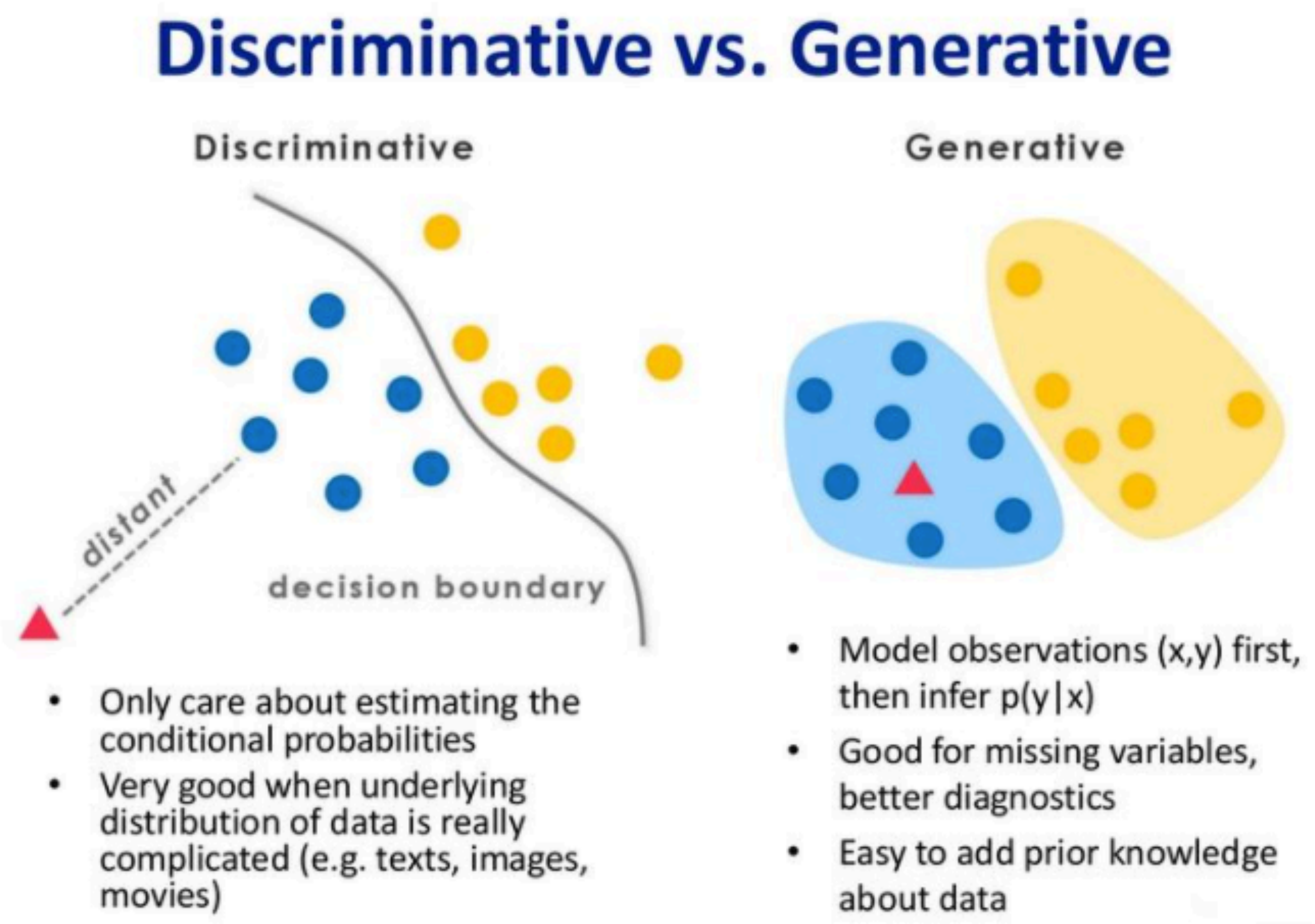
$P(X = X_i, Y = y)$

$= P(x^1 = x_i^1, x^2 = x_i^2, x^3 = x_i^3|Y = y)P(Y = y)$ （将X拆分为具体指标）

$= P(Y = y) \prod_1^3 P(x^d = x_i^d|Y = y)$(条件独立假设)

可用极大似然估计法估计相应的概率，目标是给X_i找到最大的$P(X = X_i, Y = y)$。

7. 判别模型与生成模型的区别?



判别模型之所以称为"判别"模型，是因为其根据X"判别"Y；**条件概率分布**$P(y|x)$

判别式模型举例：要确定一个羊是山羊还是绵羊，用判别模型的方法是从历史数据中学习到模型，然后通过提取这只羊的特征来预测出这只羊是山羊的概率，是绵羊的概率。

线性回归、感知机、逻辑斯蒂回归、SVM、神经网络...

生成模型之所以称为“生成”模型，是因为其预测的根据是联合概率P(X,Y)；**联合概率分布P(x, y)**

生成式模型举例：利用生成模型是根据山羊的特征首先学习出一个山羊的模型，然后根据绵羊的特征学习出一个绵羊的模型，然后从这只羊中提取特征，放到山羊模型中看概率是多少，在放到绵羊模型中看概率是多少，哪个大就是哪个。

朴素贝叶斯、隐马尔可夫模型、LDA...