

# 1. 贝叶斯判别

## 基本定理

条件概率：

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad (1)$$

全概率公式：

$$P(A) = P(A \cap B) + P(A \cap C) = P(A|B) \times P(B) + P(A|C) \times P(C) \quad (2)$$

为了要确认 $x$ 是属于 $w_1$ 类还是 $w_2$ 类，通常要看 $x$ 是来自 $w_1$ 类的概率大还是来自 $w_2$ 类的概率大。

似然函数：

$$p(w_i|x) = \frac{p(x|w_i) \cdot p(w_i)}{\sum_j p(x|w_j) \cdot p(w_j)} \quad (3)$$

似然比：

$$l_{12}(x) = \frac{p(x|w_1)}{p(x|w_2)} \quad (4)$$

判决阈值：

$$p(w_2)/p(w_1) \quad (5)$$

## 朴素贝叶斯

条件独立性假设：为何叫朴素？在特征 $x = (x_1, x_2, \dots, x_D)$ 时，朴素贝叶斯算法假设各个特征之间是相互独立的：

$$p(x_1, x_2, \dots, x_D|w) = \prod_{i=1}^D p(x_i|w) \quad (6)$$

最小风险判别：如果分类器判别 $x$ 属于 $w_j$ 类，但实际属于 $w_i$ 类，用 $L_{ij}$ 进行惩罚。观察样本指定为 $w_j$ 类的条件平均风险用 $r_j(x)$ 表示，则针对 $M$ 类的问题：

$$r_j(x) = \sum_{i=1}^M L_{ij} P(w_i|x) \quad (7)$$

通常，当 $i = j$ 时， $L_{ij} = 0$ ，因为判断是正确的没有损失。最小平均条件风险可写成：

$$r_j(x) = \frac{1}{p(x)} \sum_{i=1}^M L_{ij} P(w_i|x) P(w_i) \quad (8)$$

通常可以省略去 $p(x)$ 写为：

$$r_j(x) = \sum_{i=1}^M L_{ij} P(w_i|x) P(w_i) \quad (9)$$

通常 $L$ 可取值：

$$L_{ij} = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \quad (10)$$

则此时最小平均条件风险可表示为：

$$\begin{aligned} r_j(x) &= \sum_{i=1}^M L_{ij} P(w_i|x) P(w_i) \\ &= L_{1j} P(x|w_1) P(w_1) + L_{2j} P(x|w_2) P(w_2) + \cdots + L_{Mj} P(x|w_M) P(w_M) \\ &= \sum_{i=1}^M P(w_i|x) P(w_i) - P(w_j|x) P(w_j) \\ &= P(x) - P(w_j|x) P(w_j) \end{aligned} \quad (11)$$

这也是贝叶斯分类器，只是它的判别方法不是按错误概率最小作为标准，而是按平均条件风险作为标准。

最小平均风险条件风险分类器：对每一个 $x$ 计算出全部类别的平均风险值 $r_1(x), r_2(x), \dots, r_M(x)$ ，并将 $x$ 指定为是具有最小风险值的那一类。

## 正态分布模式的朴素贝叶斯分类器

通常这种情况是在给定的数据是连续而非离散的情况，假设在 $w_i$ 类样本的第 $i$ 个属性上取均值和方差，则有：

$$p(x|w_i) = \frac{1}{\sqrt{2\pi}\sigma_{c,i}} \exp\left(-\frac{(x - \mu_{c,i})^2}{2\sigma_{c,i}^2}\right) \quad (12)$$

已知类别 $w_i$ 的判别函数可以写成如下：

$$d_i(x) = P(x|w_i) P(w_i) \quad (13)$$

或者： $d_i(x) = \ln P(x|w_i) + \ln P(w_i)$

若两类都满足正态分布：

$$\begin{aligned} d_1(x) - d_2(x) &= \ln P(w_1) - \ln P(w_2) + (m_1 - m_2)^\top C^{-1} x - \\ &\quad \frac{1}{2} m_1^\top C^{-1} m_1 + \frac{1}{2} m_2^\top C^{-1} m_2 \end{aligned} \quad (14)$$

矩阵格式的正态类概率密度函数：

$$p(x|w_i) = \frac{1}{\sqrt{2\pi}|C_i|^{1/2}} \exp\left(-\frac{1}{2}(x - m_i)^\top C_i^{-1}(x - m_i)\right) \quad (15)$$

其中 $m_i$ 为均值向量， $C_i$ 为协方差矩阵：

$$m_i = E_i\{x\} \quad (16)$$

$$C_i = E_i\{(x - m_i)(x - m_i)^\top\} \quad (17)$$

$|C_i|$  表示  $C_i$  的行列式的值，矩阵的逆一般求法：

$$[A|I_n] \xrightarrow{\text{行变换}} [U|B] \xrightarrow{\text{行变换}} [I_n|A^{-1}] \quad (18)$$

## 朴素贝叶斯过程

首先估算概率与条件概率：

$$P(Y = c_k) = \frac{\sum_{i=1}^N I(y_i = c_k)}{N}, \quad k = 1, 2, \dots, K \quad (19)$$

离散条件：

$$P(X^{(j)} = a_{jl}|Y = c_k) = \frac{\sum_{i=1}^N I(x_i^{(j)} = a_{jl}, y_i = c_k)}{\sum_{i=1}^N I(y_i = c_k)} \quad (20)$$

连续条件：

$$p(x|w_i) = \frac{1}{\sqrt{2\pi}|C_i|^{1/2}} \exp\left(-\frac{1}{2}(x - m_i)^\top C_i^{-1}(x - m_i)\right) \quad (21)$$

对于给定的实例  $x = (x^{(1)}, x^{(2)}, \dots, x^{(n)})^\top$ ：

$$P(Y = c_k) \prod_{j=1}^n P(X^{(j)} = x^{(j)}|Y = c_k), \quad k = 1, 2, \dots, K \quad (22)$$

确定实例  $x$  的类别：

$$y = \arg \max_{c_k} P(Y = c_k) \prod_{j=1}^n P(X^{(j)} = x^{(j)}|Y = c_k) \quad (23)$$

也可以通过最小风险选择。

## Laplace 平滑

$$p(x = j) = \frac{\sum_{i=1}^N I_{x^i=j} + 1}{N + K}, \quad j = 1, \dots, K \quad (24)$$

# 2. 判别函数（分类问题）

多类情况1：  $M$  类问题，用线性判别函数将属于  $w_i$  类的模式与不属于  $w_i$  类的模式分开，需要至少  $M$  个

多类情况2：  $M$  类问题，采用每对划分，即  $w_i/w_j$  两分法，需要  $M(M - 1)/2$  个判别函数。

### 3. Fisher线性判别

出发点：降低维数

考虑把 $d$ 维空间的样本投影到一条直线上，形成一维空间。如何选择投影方向使得类别能够被分开？

假设一集合包含 $N$ 个 $d$ 维样本 $x^1, x^2, \dots, x^N$ ，其中 $N_1$ 个属于 $w_1$ 类，记为 $\Gamma_1$ ，而 $N_2$ 个属于 $w_2$ 类，记为 $\Gamma_2$ 。对 $x^n$ 做投影可得：

$$y_n = w^\top x^n, \quad n = 1, 2, \dots, N \quad (25)$$

通常 $|w|$ 的值不重要，因为只是大小，而其导致的方向是最为重要的。

基本参数：

各样本的均值：

$$m = \frac{1}{N_i} \sum_{x \in \Gamma_i} x \quad (26)$$

样本内的离散度矩阵 $S_i$ 和总样本类内离散度矩阵 $S_w$ ：

$$\begin{aligned} S_i &= \sum_{x \in \Gamma_i} (x - m_i)(x - m_i)^\top \\ S_w &= \sum_i S_i \end{aligned} \quad (27)$$

其中 $S_w$ 是半正定矩阵，当 $N > d$ 时是非奇异的。

样本类间离散度矩阵 $S_b$ ：

$$S_b = (m_1 - m_2)(m_1 - m_2)^\top \quad (28)$$

在一维空间：

各类样本的均值 $\tilde{m}_i$ ：

$$\tilde{m}_i = \frac{1}{N_i} \sum_{y \in \Gamma'_i} y \quad (29)$$

样本类内离散度 $\tilde{S}_i^2$ 和总样本类内离散度 $\tilde{S}_w$ ：

$$\begin{aligned} \tilde{S}_i^2 &= \sum_{y \in \Gamma'_i} (y - \tilde{m}_i)^2 \\ \tilde{S}_w &= \sum_i \tilde{S}_i^2 \end{aligned} \quad (30)$$

**Fisher**准则函数

$$J_F(w) = \frac{(\tilde{m}_1 - \tilde{m}_2)^2}{\tilde{S}_1^2 + \tilde{S}_2^2} \quad (31)$$

而

$$\tilde{m}_i = w^\top m_i \quad (32)$$

$$\begin{aligned}
\tilde{S}_i^2 &= \sum_{y \in \Gamma'_i} (y - \tilde{m}_i)^2 = \sum_{x \in \Gamma_i} (\mathbf{w}^\top \mathbf{x} - \mathbf{w}^\top \mathbf{m}_i)^2 \\
&= \mathbf{w}^\top \left[ \sum_{x \in \Gamma_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^\top \right] \mathbf{w} \\
&= \mathbf{w}^\top \left[ \sum_{x \in \Gamma_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^\top \right] \mathbf{w} \\
&= \mathbf{w}^\top \mathbf{S}_i \mathbf{w}
\end{aligned} \tag{33}$$

故：

$$J_F(w) = \frac{\mathbf{w}^\top \mathbf{S}_b \mathbf{w}}{\mathbf{w}^\top \mathbf{S}_w \mathbf{w}} \tag{34}$$

计算最佳变换向量  $w^*$

为了求使  $J_F(w)$  取最大值的  $w^*$ ，可采用拉格朗日乘数法进行求解。

定义拉格朗日函数： $L(w, \lambda) = \mathbf{w}^\top \mathbf{S}_b \mathbf{w} - \lambda(\mathbf{w}^\top \mathbf{S}_w \mathbf{w} - c)$

然后对  $w$  求偏导：

$$\frac{\partial L(\mathbf{w}, \lambda)}{\partial \mathbf{w}} = \mathbf{S}_b \mathbf{w} + \mathbf{S}_b^T \mathbf{w} - \lambda (\mathbf{S}_w \mathbf{w} + \mathbf{S}_w^T \mathbf{w}) = 2\mathbf{S}_b \mathbf{w} - \lambda 2\mathbf{S}_w \mathbf{w} \tag{35}$$

令偏导数为 0：

$$\mathbf{S}_b \mathbf{w}^* = \lambda \mathbf{S}_w \mathbf{w}^* \tag{36}$$

由于  $\mathbf{S}_b \mathbf{w}^* = (m_1 - m_2)(m_1 - m_2)^\top \mathbf{w}^*$ ，

而  $(m_1 - m_2)^\top \mathbf{w}^*$  为一个标量，不影响  $m_1 - m_2$  这个方向，用  $R$  表示。

故：

$$\mathbf{w}^* = \frac{R}{\lambda} \mathbf{S}_w^{-1} (m_1 - m_2) \tag{37}$$

其中  $\frac{R}{\lambda}$  取任何值都没有问题。多类问题的推导基本一致。

注：矩阵运算常用的三个求导问题：

$$\begin{aligned}
\frac{\partial A\mathbf{x}}{\partial \mathbf{x}} &= A^\top \\
\frac{\partial \mathbf{x}^\top \mathbf{x}}{\partial \mathbf{x}} &= 2\mathbf{x} \\
\frac{\partial \mathbf{x}^\top A\mathbf{x}}{\partial \mathbf{x}} &= Ax + A^\top x
\end{aligned} \tag{38}$$

## 4. 感知器算法

## 训练算法：

两类问题分别属于 $c_1$ 类和 $c_2$ 类问题，初始权重为 $w(1)$ ，若 $x^k \in c_1$ 则 $w^\top(k)x^k > 0$ ，否则若 $x^k \in c_2$ 则 $w^\top(k)x^k \leq 0$ 。

第 $k$ 步训练：

若 $x^k \in c_1$ ，而 $w^\top(k)x^k \leq 0$ ，则对第 $k$ 个模式 $x^k$ 做惩罚， $w(k+1) = w(k) + Cx^k$ ，其中 $C$ 为一个校正增量。

若 $x^k \in c_2$ ，而 $w^\top(k)x^k > 0$ ，则对第 $k$ 个模式 $x^k$ 做惩罚， $w(k+1) = w(k) - Cx^k$ ，其中 $C$ 为一个校正增量。

若分类正确，则 $w(k+1) = w(k)$

## 收敛性证明

假设模式类别是线性可分的，感知器算法可以在有限的迭代步骤里面求出权向量。

思路：第 $k+1$ 次迭代权重比第 $k$ 次更接近解矢量，则收敛。

由于线性可分，则存在 $w^*$ 使 $(w^*)^\top x > 0$

即证明 $\|w(k+1) - \alpha w^*\|^2 \leq \|w(k) - \alpha w^*\|^2$

证：

$$\begin{aligned} w(k+1) - \alpha w^* &= w(k) + x_k - \alpha w^* \\ \|w(k+1) - \alpha w^*\|^2 &= \|w(k) - \alpha w^*\|^2 + 2(w(k) - \alpha w^*)^\top x_k + \|x\|^2 \\ &\leq \|w(k) - \alpha w^*\|^2 - 2\alpha(w^*)^\top x_k + \|x\|^2 \end{aligned} \tag{39}$$

设 $\beta^2 = \max_k \|x_k\|^2$ 而 $\gamma = \min_k (w^*)^\top x_k > 0$ ，则：

$$\|w(k+1) - \alpha w^*\|^2 < \|w(k) - \alpha w^*\|^2 - 2\alpha\gamma + \beta^2 \tag{40}$$

若 $\alpha = \beta^2/\gamma$ 则：

$$\|w(k+1) - \alpha w^*\|^2 < \|w(k) - \alpha w^*\|^2 - \beta^2 \tag{41}$$

经过 $k$ 步后，权系数：

$$\|w(k+1) - \alpha w^*\|^2 < \|w(1) - \alpha w^*\|^2 - k\beta^2 \tag{42}$$

由于不能为负数，故经过不超过 $k_0 = \frac{\|w(1) - \alpha w^*\|^2}{\beta^2}$ 步即终止。

## 多类训练算法

存在 $M$ 类的判别函数 $\{d_i, i = 1, 2, \dots, M\}$ ，若 $x_k \in c_i$ ，则 $d_i > d_j, \forall j \neq i$ 。

设有 $M$ 个模式类别 $c_1, c_2, \dots, c_M$ ，则在第 $k$ 次迭代时，若一个属于 $c_i$ 的模式样本 $x$ ，先计算 $M$ 个模式的判别函数：

$$d_j(k) = w_j(k)x, j = 1, 2, \dots, M$$

若第 $j$ 个权向量使 $d_i(k) < d_j(k)$ ，则调整权向量：

$$w_i(k+1) = w_i(k) + Cx$$

$$w_j(k+1) = w_j(k) - Cx$$

其他权向量的数值保持不变。

如果分类正确也保持权向量不变  $w_j(k+1) = w_j(k)$ 。

## 5. 势函数

可能不考

## 6. 特征选择及K-L变换

### 特征选择准则

特征选择目的主要是为了在保留识别信息的前提下，降低特征空间的维度。

类别可分性准则应具有如下特点：不同类别模式特征的均值向量之间的距离应该最大，而属于同一类的模式特征其方差之和最小。

若原始特征测量量值是独立统计的，此时只需要对样本独立进行分析，从中选择m个最好的作为分类特征。

一般特征的散布矩阵准则：

- 类内、类间散布矩阵  $S_w$  和  $S_b$ 。类间离散度越大且类内离散度越小最好。
- 散布矩阵准则  $J_1$  和  $J_2$  形式  $J_1 = \det S_w^{-1} S_b = \prod_i \lambda_i$ ,  $J_2 = \text{tr } S_w^{-1} S_b = \sum_i \lambda_i$ , 找到使  $J_1$  或  $J_2$  最大的子集作为所选择的分类特征。

### K-L变换

适用于任意概率密度函数的正交变换。

离散K-L变换展开：

设有一连续的随机实函数  $x(t)$ ，其中  $T_1 \leq t \leq T_2$ ，则  $x(t)$  可用正交函数集  $\{\varphi_j(i), j = 1, 2, \dots\}$  的线性组合展开：

$$\begin{aligned} x(t) &= a_1 \varphi_1(t) + a_2 \varphi_2(t) + \dots + a_j \varphi_j(t) + \dots \\ &= \sum_{j=1}^{\infty} a_j \varphi_j(t), \quad T_1 \leq t \leq T_2 \end{aligned} \tag{43}$$

其中  $a_j$  为展开式的随机系数， $\varphi_j(t)$  为一组连续的正交函数，满足：

$$\int_{T_1}^{T_2} \varphi_n^{(t)} \tilde{\varphi}_m(t) dt = \begin{cases} 1, & m = n \\ 0, & m \neq n \end{cases} \quad (44)$$

其中  $\tilde{\varphi}_m(t)$  为  $\varphi_n^{(t)}$  的共轭复数式。

写成离散形式的正交函数形式，使连续随机函数  $x(t)$  和连续正交函数  $\varphi_j(t)$  被等间隔采样为  $n$  个离散点，即：

$$\begin{aligned} x(t) &\rightarrow \{x(1), x(2), \dots, x(n)\} \\ \varphi_j(t) &\rightarrow \{\varphi_j(1), \varphi_j(2), \dots, \varphi_j(n)\} \end{aligned} \quad (45)$$

向量：

$$\begin{aligned} x &= (x(1), x(2), \dots, x(n))^\top \\ \varphi_j &= (\varphi_j(1), \varphi_j(2), \dots, \varphi_j(n)), \quad j = 1, 2, \dots, n \end{aligned} \quad (46)$$

写为离散展开式：

$$x = \sum_{j=1}^n a_j \varphi_j = \Phi a, \quad T_1 \leq t \leq T_2 \quad (47)$$

其中， $a$  为展开式的随机系数的向量形式：

$$a = (a_1, a_2, \dots, a_j, \dots, a_n)^\top \quad (48)$$

$\Phi$  为  $n \times n$  的矩阵，即：

$$\Phi = (\varphi_1, \varphi_2, \dots, \varphi_n) = \begin{bmatrix} \varphi_1(1) & \varphi_2(1) & \cdots & \varphi_n(1) \\ \varphi_1(2) & \varphi_2(2) & \cdots & \varphi_n(2) \\ \cdots & \cdots & \cdots & \cdots \\ \varphi_1(n) & \varphi_2(n) & \cdots & \varphi_n(n) \end{bmatrix} \quad (49)$$

$\Phi$  将  $x$  变换为  $a$ 。

**K-L** 展开式系数计算过程：

1. 给定一系列随机向量（样本） $x$ ，首先保证  $E[x] = 0$ ，否则对其归一化，就是减均值。
2. 求随机向量的自相关矩阵： $R = E\{xx^\top\}$
3. 然后，求出矩阵  $R$  的特征值  $\lambda_j$  与特征向量  $\varphi_j$ ,  $j = 1, 2, \dots, n$ ，这样可以满足上述的正交函数形式，然后列出矩阵  $\Phi = (\varphi_1, \varphi_2, \dots, \varphi_n)$
4. 计算变换后的展开式： $a = \Phi^\top x$

**K-L** 特征选择问题证明：

注：可以不选  $n$  个，也可以只选  $m$  个， $m < n$ ，选择特征值最大的前  $m$  个特征。

$$\Phi = (\varphi_1, \varphi_2, \dots, \varphi_m) \quad (50)$$

由于我们在选择了变换矩阵后，我们要使得降维后的向量在最小均方差条件下接近原来的向量  $x$ ，对于  $x = \sum_{j=1}^n a_j \varphi_j$ ，现只取  $m$  项，对于省略部分用常数  $b$  代替，则：

$$\hat{x} = \sum_{j=1}^m a_j \varphi_j + \sum_{j=m+1}^n b \varphi_j \quad (51)$$

误差：

$$\Delta x = x - \hat{x} = \sum_{j=m+1}^n (a_j - b) \varphi_j \quad (52)$$

均方误差:

$$E\{\|\Delta x\|\}^2 = \sum_{j=m+1}^n E(a_j - b)^2 \quad (53)$$

为了让误差最小, 对应**b**应满足:

$$\frac{\partial}{\partial b}[E(a_j - b)^2] = \frac{\partial}{\partial b}[E(a_j^2 - 2a_j b + b^2)] = 2[E(a_j) - b] = 0 \quad (54)$$

则应该使  $b = E[a_j]$ , 即对省略了的  $a$  的分量, 此时应该满足:

$$\begin{aligned} E\{\|\Delta x\|\}^2 &= \sum_{j=m+1}^n E(a_j - E[a_j])^2 = \sum_{j=m+1}^n [\varphi_j^\top(x - E[x])(x - E[x])^\top \varphi_j] \\ &= \sum_{j=m+1}^n \varphi_j^\top C_x \varphi_j \end{aligned} \quad (55)$$

设  $\lambda_j$  为  $C_x$  的第  $j$  个特征值 (因为我们计算特征值时候就是基于这个协方差矩阵或者自相关矩阵计算的),  $\varphi_j$  为  $\lambda_j$  对应的特征向量, 则:

$$C_x \varphi_j = \lambda_j \varphi_j \quad (56)$$

由于  $\varphi_j^\top \varphi_j = 1$  故  $\varphi_j^\top C_x \varphi_j = \lambda_j$

因此:

$$E\{\|\Delta x\|\}^2 = \sum_{j=m+1}^n \varphi_j^\top C_x \varphi_j = \sum_{j=m+1}^n \lambda_j \quad (57)$$

故被遗弃的特征的特征值应越小越好。

**K-L变换总结:**

K-L变换前需要将  $E[x] = 0$ , 因此应先将均值作为新坐标轴的原点, 然后再采用协方差矩阵和自相关矩阵计算特征值。

采用K-L变换作为模式分类的特征提取时, 需要特别注意保留不同类别的模式分类鉴别信息, 仅单纯考虑尽可能代表原来模式的主成分, 有时并不一定有利于分类的鉴别。

## 7.逻辑回归LR

最小二乘法:

最优化问题:  $\min_w J(w) = \sum_{i=1}^N (w^\top x^i - y^i)^2$

梯度下降:  $\frac{\partial J(w)}{\partial w} = 2 \sum_{i=1}^N x_j^i (w^\top x^i - y^i)$

更新规则:

批量梯度下降:  $w_j = w_j - 2\alpha \sum_{i=1}^N x_j^i (w^\top x^i - y^i)$ ,  $\alpha > 0$

随机梯度下降:  $w_j = w_j - 2\alpha x_j^i (w^\top x^i - y^i)$ ,  $\alpha > 0$

判别式模型, 二分类逻辑回归:

估计后验概率  $p(y|x)$

$$P(y=1|x) = f(x, w) = \text{Sigmoid}(w^\top x) = \frac{1}{1 + \exp -w^\top x} \quad (58)$$

概率分布 (伯努利分布) :

$$P(y|x, w) = (f(x, w))^y (1 - f(x, w))^{1-y} \quad (59)$$

似然:

$$L(w) = \prod_{i=1}^N P(y^i|x^i, w) = \prod_{i=1}^N (f(x^i, w))^{y^i} (1 - f(x^i, w))^{1-y^i} \quad (60)$$

最大化 log 似然:

$$l(w) = \log L(w) = \sum_{i=1}^N (y^i \log f(x^i, w) + (1 - y^i) \log (1 - f(x^i, w))) \quad (61)$$

梯度:

$$\frac{\partial l(w)}{\partial w_j} = (y^i - f(x^i, w)) x_j^i \quad (62)$$

SGD:

$$w_j = w_j + \alpha (y^i - f(x^i, w)) x_j^i \quad (63)$$

注: sigmoid 函数  $f(x)$  求导:  $f(x)(1 - f(x))$ , log 函数求导:  $\frac{d \log_a x}{dx} = \frac{1}{x \ln a}$ 。

多分类逻辑回归:

这里用 softmax 代替 sigmoid

$$P(C_k|x, w) = \frac{\exp(w_k^\top x)}{\sum_{j=1}^K \exp(w_j^\top x)} \quad (64)$$

概率分布:

$$\begin{aligned} \mu_i &= P(y_i = 1|x, w) \\ P(y, \mu) &= \prod_{i=1}^K \mu_i^{y_i} \end{aligned} \quad (65)$$

则:

$$\begin{aligned} P(y^1, y^2, \dots, y^N | w_1, w_2, \dots, w_k, x^1, x^2, \dots, x^N) \\ = \prod_{i=1}^N \prod_{j=1}^K P(C_k | w_k, x^i) \end{aligned} \quad (66)$$

优化 (交叉熵损失函数) :

$$\begin{aligned} \min E &= -\ln P(y^1, y^2, \dots, y^N | w_1, w_2, \dots, w_k, x^1, x^2, \dots, x^N) \\ &\quad \min - \sum_{i=1}^N \sum_{k=1}^K y_k^i \ln \mu_{ik} \end{aligned} \tag{67}$$

梯度：

$$\begin{aligned} \nabla_{w_j} E &= \sum_{i=1}^N \left( \sum_{k=1}^K y_k^i \frac{\mu'_{ij}}{\mu_{ij}} \right) \\ &= \sum_{i=1}^N (\mu_{ij} - y_j^i) x^i \end{aligned} \tag{68}$$

## LR (逻辑回归) 与NB (朴素贝叶斯)

当模型假设正确，NB和LR产生相似的分类器

假设不争取时候，LR偏差较小，预期LR优于NB

NB收敛速度一般比LR快

## 8.MLE与MAP

### MLP

假设给定函数，输入给了高斯噪声 $\varepsilon$ ，则 $y = f(x, w) + \varepsilon$ ，均值是0，方差为 $\beta^{-1}$ 的高斯噪声：

似然函数：

$\prod_{i=1}^N N(y^i | w^\top x^i, \beta^{-1})$ 满正太分布。

其对数似然：

$$\begin{aligned} &\sum_{i=1}^N \ln N(y^i | y^i | w^\top x^i, \beta^{-1}) \\ &= \sum_{i=1}^N \ln \left[ \frac{1}{\sqrt{2\pi}\sqrt{\beta^{-1}}} \exp \left( -\frac{(f(x^i, w) - y^i)^2}{2\beta^{-1}} \right) \right] \\ &= \sum_{i=1}^N \left[ \frac{1}{2} \ln \beta - \frac{1}{2} \ln 2\pi - \frac{1}{2} \beta (f(x^i, w) - y^i)^2 \right] \\ &= \frac{N}{2} \ln \beta - \frac{N}{2} \ln 2\pi - \frac{1}{2} \beta \sum_{i=1}^N (f(x^i, w) - y^i)^2 \\ &= \frac{N}{2} \ln \beta - \frac{N}{2} \ln 2\pi - \frac{1}{2} \beta J(w) \end{aligned} \tag{69}$$

相当于最小化平方误差之和。

## MAP

---

贝叶斯:  $p(w|y) = p(y|w)p(w)/p(y)$

似然:  $p(y|X, w, \beta) = \prod_{i=1}^N N(y^i | w^\top x^i, \beta^{-1})$

先验:  $p(w) = N(0, \lambda^{-1}I)$

后验概率:

$$\ln(p(w|y)) = -\beta \sum_{i=1}^N (y^i - w^\top x^i)^2 - \lambda w^\top w + \text{constant} \quad (70)$$

最大化后验概率, 等同于最小化带有正则项的平方和误差:

$$\min_w \sum_{i=1}^N (w^\top x^i - y^i)^2 + \alpha w^\top w, \quad \alpha = \frac{\lambda}{\beta} \quad (71)$$

## 总结

---

1. MLE:  $\hat{\theta}_{MLE} = \arg \max_{\theta} P(D|\theta)$
2. MAP:  $\hat{\theta}_{MAP} = \arg \max_{\theta} P(\theta|D) = \arg \max_{\theta} P(D|\theta)P(\theta)$

就是最大似然估计其实不需要知道先验  $P(\theta)$ , 但是最大后验估计是需要知道先验  $P(\theta)$

# 9. 高斯判别分析

伯努利分布:  $x \in \{0, 1\}$  的分布由连续参数  $\beta \in [0, 1]$  控制:

$$P(x|\beta) = \beta^x (1-\beta)^{1-x} \quad (72)$$

而给出  $N$  个服从伯努利分布的样本中, 观察到  $m$  次  $x=1$  的概率:

$$P(m|N, \beta) = C_N^m \cdot \beta^m (1-\beta)^{N-m} \quad (73)$$

多项式分布, 即取  $K$  个状态, 第  $k$  个状态被观测到  $m_k$  次的概率:

$$P(m_1, \dots, m_K|N, \beta) = \binom{N}{m_1, \dots, m_K} \prod_{k=1}^K \beta_k^{m_k} \quad (74)$$

GDA 多变量正态分布建模:

$$y \sim Bernoulli(\beta); x|y=0 \sim N(\mu_0, \Sigma); x|y=1 \sim N(\mu_1, \Sigma) \quad (75)$$

Log似然:

$$\begin{aligned} L(\beta, \mu_0, \mu_1, \Sigma) &= \log \prod_{i=1}^N p(x^i, y^i; \beta, \mu_0, \mu_1, \Sigma) \\ &= \log \prod_{i=1}^N p(y^i | \beta) p(x^i | y^i; \beta, \mu_0, \mu_1, \Sigma) \end{aligned} \quad (76)$$

MLE:

$$\beta = \frac{1}{N} \sum_{i=1}^N I_{y^i=1}; \mu_k = \frac{\sum_{i=1}^N I_{y^i=k} x_i}{\sum_{i=1}^N I_{y^i=k}}, k = \{0, 1\}; \Sigma = \frac{1}{N} \sum_{i=1}^N (x^i - \mu_{y^i})(x^i - \mu_{y^i})^\top \quad (77)$$

总结:

GDA有很强的模型假设，当假设正确时，处理数据的效率更高

LR假设很弱，因此对偏离假设时具有鲁棒性

实际中LR更常用

## 10.SVM

函数间隔：给定一个样本 $(x_i, y_i)$ ，它到 $(w, b)$ 确定的超平面的函数间隔为：

$$\hat{\gamma}_i = y_i(w^\top x_i + b) \quad (78)$$

则定义超平面 $(w, b)$ 的所有样本点 $(x_i, y_i)$ 的函数间隔最小值：

$$\hat{\gamma} = \min_{i=1, \dots, N} \hat{\gamma}_i \quad (79)$$

几何间隔：为了使函数间隔具有不变性

$$\gamma_i = y_i \left( \frac{w}{\|w\|} \cdot x_i + \frac{b}{\|w\|} \right) \quad (80)$$

样本中的几何间隔最小值：

$$\gamma = \min_{i=1, \dots, N} \gamma_i \quad (81)$$

几何间隔与函数间隔的关系：

$$\gamma = \frac{\hat{\gamma}}{\|w\|} \quad (82)$$

间隔最大化：

$$\max_{\gamma} \frac{\gamma}{\|w\|} \quad (83)$$

约束条件：

$$\gamma(w^\top x + b) \geq \gamma \quad (84)$$

硬间隔中，令 $\gamma = 1$

$$\text{优化: } \max_{\gamma} \frac{\gamma}{\|w\|} = \min_w \frac{1}{2} \|w\|^2$$

$$\text{约束: } y(w^\top x + b) \geq 1$$

对偶算法:

$$L(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^N \alpha_i y_i (w^\top x_i + b) + \sum_{i=1}^N \alpha_i \quad (85)$$

本质是在优化:

$$\max_{\alpha} \min_{w,b} L(w, b, \alpha) \quad (86)$$

先对 $w, b$ 取极小:

$$\begin{aligned} & \min_{w,b} L(w, b, \alpha) \\ & \nabla_w L = w - \sum_{i=1}^N \alpha_i y_i x_i \\ & \nabla_b L = - \sum_{i=1}^N \alpha_i y_i \end{aligned} \quad (87)$$

设最优为 $w^*$ :

$$\begin{aligned} w^* &= \sum_{i=1}^N \alpha_i y_i x_i \\ \sum_{i=1}^N \alpha_i y_i &= 0 \\ \alpha_i &\geq 0 \end{aligned} \quad (88)$$

此时:

$$\begin{aligned} L(w, b, \alpha) &= \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j x_i x_j - \sum_{i=1}^N \alpha_i y_i \left( \sum_{j=1}^N \alpha_j y_j x_j + b \right) + \sum_{i=1}^N \alpha_i \\ &= -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j x_i x_j + \sum_{i=1}^N \alpha_i \end{aligned} \quad (89)$$

即优化:

$$\begin{aligned} & \max_{\alpha} -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j x_i x_j + \sum_{i=1}^N \alpha_i \\ & \longrightarrow \min_{\alpha} \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j x_i x_j - \sum_{i=1}^N \alpha_i \\ & s.t. \sum_{i=1}^N \alpha_i = 0, \alpha_i \geq 0 \end{aligned} \quad (90)$$

之后围绕着这个方程求解即可，可能比较麻烦，需要启发式计算。

# 11.K-means聚类

## 算法

问题：给定 $N$ 个样本点 $X = \{x_i\}_{i=1}^N$ 进行聚类

输入：数据 $\mathcal{D} = \{x_1, x_2, \dots, x_N\}$ , 聚类簇数目为 $K$ 。

随机选择 $K$ 个种子数据点作为 $K$ 个簇的中心

repeat

for each  $x_i \in \mathcal{D}$  do

计算 $x_i$ 与每一个簇中心的距离 $\text{dist}(x_i, \mu_{k'})$

将 $x_i$ 指配到距离簇中心最近的簇中心： $z_i = \arg \min_{k'} \text{dist}(x_i, \mu_{k'})$

end for

用当前的簇内点，重新计算 $K$ 个簇中心的位置

until当前簇中心未更新

备注：

欧式距离作为距离度量

每个节点都划分到距离最近的那个簇中心，用 $r_{j,k} \in \{0, 1\}$ 表示。

目标函数：

$$J = \sum_{i=1}^N \sum_{k=1}^K r_{i,k} \|x_i - \mu_k\|^2 \quad (91)$$

优化：

固定 $\mu_k$ , 优化 $r_{i,k}$

固定 $r_{i,k}$ , 优化 $\mu_k$ , 就是求为 $r_{i,k}$ 所有样本点的均值

## $K$ 的选择问题

寻找一个 $K$ , 使 $J_K - J_{K+1}$ 的改进很小。

$$K_\alpha = \min \left\{ k : \frac{J_k - J_{k+1}}{\sigma^2} \leq \alpha \right\} \quad (92)$$

其中 $\sigma^2 = \mathbb{E}[\|X - \mu\|^2]$ ,  $\mu = \mathbb{E}[X]$

## 其他聚类方法 (GMM, 层次聚类, DBSCAN)

### GMM:

假设K个簇，每个簇服从高斯分布，以概率 $\pi_k$ 随机选择一个簇，然后观测数据。

概率密度函数：

$$p(x) = \sum_{k=1}^K \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \quad (93)$$

$\pi_k$ 为混合比例， $\mathcal{N}(x|\mu_k, \Sigma_k)$ 为混合的成分。

参数估计：极大似然估计+EM算法，E步求期望，M步重新估计参数

与K-means异同：Kmeans采用虽小平方距离和损失函数，而GMM是最小化负对数的似然函数。Kmeans是硬划分到簇，而GMM是软划分。Kmeans样本属于每个簇概率相同，球形簇，而GMM簇概率不同，可以被用于椭球形簇。

### 层次聚类：

不需要提前假定簇的数目，选择树状图某一层就可以获取任意簇数量的聚类结构。

自底向上：递归合并相似度最高/距离最近的两个簇

自顶向下：递归分裂不一致的簇，例如拥有最大直径的簇

### DBSCAN，基于密度的聚类：

基于密度的聚类，将临近的密度高的区域连成一片形成簇

优点：各种大小或形状的簇，且具有一定的抗噪声能力

优势：不需要簇的数目，可以对任意形状簇进行聚类，对离群点比较鲁棒

劣势：参数选择困难，不适合密度差异大的数据集，计算复杂度比较大。

# 12. 降维

矩阵的秩：最大线性无关行或列的数目

矩阵的迹：方阵 $A$ 的 $tr(A)$ 是对角线元素之和

奇异值分解SVD：

输入 $A_{M \times N}$ 矩阵，SVD:  $A = U\Sigma V^\top$

其中 $\Sigma = \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_R \end{bmatrix}$  奇异值， $R = \min(M, N)$

$u_k$ 、 $v_k^\top$ : 奇异值 $\sigma_k$ 对应的奇异向量

$U^\top U = I, V^\top V = I$ :  $U$ 和 $V$ 是正交矩阵

## PCA

---

目标: 最小平方重建误差和:

$$\min_{W \in \mathbb{R}^{D \times D'}} \sum_{i=1}^N \|x_i - Wz_i\|^2 \quad (94)$$

目标函数1推导:

$$\begin{aligned} & \sum_{i=1}^N \|x_i - Wz_i\|^2 \\ &= \sum_{i=1}^N \left\| \sum_{j=1}^{D'} z_{i,j} w_j - x_i \right\|_2^2 \\ &= \sum_{i=1}^N z_i^\top z_i - 2 \sum_{i=1}^N z_i^\top W^\top x_i + \text{constant} \\ &\quad \because z_i = W^\top x_i \\ &\therefore \sum_{i=1}^N \|x_i - Wz_i\|^2 \propto -\text{tr}(W^\top (\sum_{i=1}^N x_i x_i^\top) W) \\ &= -\text{tr}(W^\top X X^\top W) = -\text{tr}(Z Z^\top) = -\sum_{i=1}^N z_i^\top z_i \end{aligned} \quad (95)$$

即寻找一个方向向量, 使数据投影到其上方后方差最大。

最大化样本的投影误差:

$$\max \sum_i z_i^\top z_i = \max_W \text{tr}(W^\top X X^\top W) = \max_W \text{tr} W^\top S W \quad (96)$$

寻找 $W \in \mathbb{R}^{D \times D'}$ 使上述投影误差最大,  $W^\top W = I$

由于带约束的优化, 可以用拉格朗日乘子:

针对第一个分量 $w_1$ :

$$\begin{aligned} L &= w_1^\top S w_1 - \lambda_1 (w_1^\top w_1 - 1) \\ \frac{\partial L}{\partial w_1} &= 2S w_1 - 2\lambda_1 w_1 = 0, \\ S w_1 &= \lambda_1 w_1 \\ J &= w_1^\top S w_1 = w_1^\top \lambda_1 w_1 = \lambda_1 \end{aligned} \quad (97)$$

而 $\lambda_1$ 是最大特征值。

接着计算第二个方向投影, 约束为:  $w_2^\top w_2 = 1, w_2^\top w_1 = 0$

拉格朗日函数:  $L = w_1^\top S w_1 + w_2^\top S w_2 - \lambda_2 (w_2^\top w_2 - 1) - \lambda_{2,1} w_2^\top w_1$

$$\frac{\partial L}{\partial w_2} = 2Sw_2 - 2\lambda_2 w_2 - \lambda_{2,1}w_1 = 0$$

同时乘  $w_1^\top$ :

$$2w_1^\top Sw_2 - 2\lambda_2 w_1^\top w_2 - \lambda_{2,1}w_1^\top w_1 = 0, \text{ 由于 } w_1^\top w_1 = 1 \text{ 所以 } \lambda_{2,1} = 0$$

$Sw_2 = \lambda_2 w_2$  且  $\lambda_2$  为第二大特征值

此时  $L = \lambda_1 + \lambda_2$

然后以此类推

整体算法:

输入:  $X = (x_1, x_2, \dots, x_N)$

过程:

$$x_i = x_i - \frac{1}{N} \sum_{j=1}^N x_j$$

计算  $S = XX^\top$

对  $S$  做特征分解

$D'$  最大特征值对应的特征向量  $w_1, w_2, \dots, w_{D'}$

输出  $W = (w_1, w_2, \dots, w_{D'})$

## 流形学习: 结构保持 / 距离保持

度量型 MDS 的算法:

将  $n$  个  $D$  维原始数据降维到  $D'$

1. 计算样本的距离矩阵  $D = [d_{i,j}] = [dist(x_i, x_j)]$ , 其中  $D \in \mathbb{R}^{N \times N}$
2. 中心矫正方法构造矩阵  $B = JDJ$ ,  $J = I - \frac{1}{n}O$ , 其中  $I$  为  $n \times n$  的单位阵,  $O$  为  $n \times n$  的值全为 1 的矩阵。
3. 计算矩阵  $B$  的特征向量  $e_1, e_2, \dots, e_m$  以及对应的特征值  $\lambda_1, \lambda_2, \dots, \lambda_m$
4. 确定维度  $D'$  重构数据  $Z = E\Lambda^{1/2}$ ,  $Z \in \mathbb{R}^{D' \times N}$ , 其中  $E$  为  $D'$  个最大特征值组成的特征向量, 而  $\Lambda$  为  $D'$  个特征值构成的对角矩阵。

测地距离:

- 邻近的点: 输入空间的欧氏距离提供一个测地线距离的近似。
- 较远的点: 测点距离能够通过一系列邻域点之间的欧式距离的累加近似。

全局距离保持 ISOMAP 算法:

- 构建邻接图
- 计算最短路径:  $\min(d_G(i, j), d_G(i, k) + d_G(k, j))$ , 通常用 Floyd 算法 (复杂度  $O(N^3)$ ) 和 Dijkstra 算法 (复杂度  $O(KN^2 \log N)$ )。
- 构建低维嵌入: 保持点对之间的测地距离, 用 MDS 方法。

局部距离保持:

拉普拉斯特特征映射(Laplacian Eigenmaps):

- 令  $x_1, x_2, \dots, x_N \in \mathbb{R}^D$
- 构造相似性图，表示节点之间的邻接关系： $\varepsilon$ 邻域， $K$ 邻域
- 通过对图的拉普拉斯矩阵进行特征值分解，得到映射。

局部优先，兼顾全局：

T-SNE:

SNE将欧氏距离转换为用概率表示的相似度，原始空间中的相似度由高斯联合概率表示，嵌入空间的相似度由“学生T分布”表示。

T-SNE的目标是对所有数据点对的概率  $p_{ij}$  与  $q_{ij}$ ，最小化KL散度  $C = KL(p\|q) = \sum_i \sum_j p_{i,j} \log \frac{p_{i,j}}{q_{i,j}}$ 。  
 $p_{i|j} = \frac{\exp(-\|x_j - x_i\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_k - x_i\|^2 / 2\sigma_i^2)}$ ， $p_{i,j} = \frac{p_{i|j} + p_{j|i}}{2N}$ ，其中  $\sigma_i$  参数是以数据点  $x_i$  为中心的高斯分布的标准差，但是实际很难评估，故一般用困惑度表示，原始论文困惑度建议设为5-50，为固定数。

优化可用随机梯度下降法。

为什么用T分布？T分布与正太分布很类似，中间比正态分布低，但是两侧比正态分布高。当高维空间中距离较小的点对，概率一致时候，T分布的在低维空间距离更小，相似样本更紧致，距离较大的点对距离更大，不相似样本更远。

T-SNE缺点：计算复杂， $O(N^2)$ ；全局结构没有明确保留；主要用于可视化。

UMAP:

步骤：学习高维空间中的流形结构，找到该流形的低维表示。

- 寻找最近邻，例如KNN算法
- 根据K近邻，构建相似图（变化距离，局部连接，概率&边的权重  $p_{j|i}$ ，合并边的权重  $p_{i,j}$ ）
- 优化，目标函数为交叉熵：

$$J = \sum_{e_{i,j} \in \varepsilon} p_{i,j} \log \frac{p_{i,j}}{q_{i,j}} + (1 - p_{i,j}) \log \frac{(1 - p_{i,j})}{(1 - q_{i,j})} \quad (98)$$

优点：不仅可以可视化，还可以降维；更快，用随机梯度下降法；更好全局结构

## 13. 集成学习

无免费午餐定理：模型的选取要以问题的特点为根据。

丑小鸭定理：\*\*

奥卡姆剃刀：在性能相同的情况下，应该选取更加简单的模型。

过于简单的模型会导致欠拟合，过于复杂的模型会导致过拟合。

从误差分解的角度看，欠拟合模型的偏差较大，过拟合模型的方差较大。

# Bagging原理及降低Variance推导

对给定的 $N$ 个样本的数据集 $\mathcal{D}$ 进行Bootstrap采样，得到 $\mathcal{D}^1$ ，在 $\mathcal{D}^1$ 上训练模型 $f_1$

上述过程重复 $M$ 次，得到 $M$ 个模型，则 $M$ 个模型的平均回归/投票为：

$$f_{avg}(x) = \frac{1}{M} \sum_{m=1}^M f_m(x) \quad (99)$$

bootstrap样本：通过对原始的 $N$ 个样本数据 $\mathcal{D} = \{x_1, \dots, x_N\}$ 进行 $N$ 次有放回的采样 $N$ 个数据 $\mathcal{D}'$ ，则称为一个bootstrap样本。

对原始数据进行有放回的随机采样，抽取的样本数目同原始样本数目一样。

一个样本不在采样集中概率： $(1 - \frac{1}{N})^N$ ，约有63.2%的样本出现在采样集中。

为什么Bagging可降低模型的方差：

令随机变量 $X$ 的均值为 $\mu$ ，方差为 $\sigma^2$

则 $N$ 个独立同分布的样本的样本均值 $\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i$

则样本均值 $\bar{X}$ 的期望为： $\mathbb{E}(\bar{X}) = \frac{1}{N} \sum_{i=1}^N \mathbb{E}(X_i) = \mu$ ，故 $\bar{X}$ 是 $X$ 的无偏估计。

而样本均值 $\bar{X}$ 的方差为： $Var(\bar{X}) = \frac{1}{N^2} \sum_{i=1}^N Var(X_i) = \frac{\sigma^2}{N}$

# Boosting推导

Boosting是弱学习器按顺序学习的，第 $n-1$ 个学习器 $\phi_{n-1}(x)$ 要能帮助第 $n$ 个学习器 $\phi_n(x)$ ，然后组合所有的弱学习器 $f(x) = \sum_{m=1}^M \alpha_m \phi_m(x)$ 。

分对的样本减少权重，分错的样本增加权重。

算法描述：

给定训练集： $(x_1, y_1), \dots, (x_n, y_n)$ ，其中 $y_i \in \{-1, 1\}$ 表示 $x_i$ 的类别标签

对 $m = 1 : M$ ：

对训练样本采用权重 $w_{m,i}$ ，计算弱分类器 $\phi_m(x)$

计算该弱分类器 $\phi_m(x)$ 在分布 $w_m$ 上的误差： $\varepsilon_m = \sum_{i=1}^N w_{m,i} \mathbb{I}(\phi_m(x_i) \neq y_i)$

计算 $d_m = \sqrt{(1 - \varepsilon_m)/\varepsilon_m}$ ,  $\alpha = \log d_m = \frac{1}{2} \log \frac{1-\varepsilon_m}{\varepsilon_m}$

更新训练样本的分布： $w_{m+1,i} = \begin{cases} \frac{w_{m,i}/d_m}{z_m} & y_i \phi_m(x_i) = 1 \\ \frac{w_{m,i}d_m}{z_m} & y_i \phi_m(x_i) = -1 \end{cases}$

$= \frac{w_{m,i}d_m^{-y_i\phi_m(x_i)}}{Z_m} = \frac{w_{m,i} \exp(-\alpha_m y_i \phi_m(x_i))}{Z_m}$ ，其中 $Z_m$ 为归一化常数，使 $w_{m+1}$ 是个分布。

强分类器： $f(x) = sgn(\sum_{m=1}^M \alpha_m \phi_m(x))$

证明减少偏差：

$$\therefore w_{M+1,i} = w_{M,i} \frac{\exp(-\alpha_M y_i \phi_M(x_i))}{Z_M} = w_{1,i} \frac{\exp\left(-y_i \sum_{m=1}^M \alpha_m \phi_m(x)\right)}{\prod_{m=1}^M Z_m} \quad (100)$$

其中  $\sum_{i=1}^N w_{M+1,i} = 1$ , 所以  $\prod_{m=1}^M Z_m = \sum_{i=1}^N w_{1,i} \exp(-y_i f(x_i)) = \frac{1}{N} \sum_{i=1}^N \exp(-y_i f(x_i))$ 。

训练误差：

$$ERR_{train}(f(x)) = \frac{1}{N} \sum_{i=1}^N \exp(-y_i f(x_i)) \quad (101)$$

而：

$$\begin{aligned} Z_m &= \sum_{i=1}^N w_{m,i} = \sum_{i=1}^N w_{m,i} d_m \mathbb{I}(y_i \neq \phi_m(x_i)) + \sum_{i=1}^N w_{m,i} / d_m \mathbb{I}(y_i = \phi_m(x_i)) \\ &= d_m \varepsilon_m + 1/d_m (1 - \varepsilon_m) \\ \therefore d_m &= \sqrt{(1 - \varepsilon_m)/\varepsilon_m} \\ . &:= 2\sqrt{(1 - \varepsilon_m)\varepsilon_m} \end{aligned} \quad (102)$$

训练误差：

$$\begin{aligned} ERR_{train}(f(x)) &= \frac{1}{N} \sum_{i=1}^N \exp(-y_i f(x_i)) \\ &= \prod_{m=1}^M Z_m = 2^M \prod_{m=1}^M \sqrt{(1 - \varepsilon_m)\varepsilon_m} \end{aligned} \quad (103)$$

由于  $0 < \varepsilon_m < 0.5, 0 < 1 - \varepsilon_m < 1$

所以误差  $ERR_{train}$  越来越小。

## 14. 半监督学习

基本假设：

聚类假设：如果两个点都在同一个簇，那么很可能属于同一个类别。

低密度分割：决策边界应该在低密度区域

平滑假设：如果高密度区域有两个点  $x_1, x_2$  距离较近，那么对应的输出  $y_1, y_2$  也应该比较近

流形假设：高维数据大致会分布在一个低维的流形上，流形上临近的样本拥有相似的输出（平滑性假设相似）

自学习算法：

- 假设：输出高置信度的预测是正确的
- 缺点：早期的错误会被强化，收敛方面没有保障。

多视角学习（协同学习）：

训练两个分类器 $f_1, f_2$ ，他们针对的特征不同（特征分割），把 $f_1$ 的 $k$ 个最高置信结果给 $f_2$ ， $f_2$ 的 $k$ 个最高置信结果给 $f_1$ 做标签。

优点：对于错误没有那么敏感，且可以用到已有的各种分类器

缺点：自然的特征分裂不存在，使用全部特征的模型可能效果更好。

本质是搜索最好的分类器

生成式半监督模型，GMM：

GMM使用MLE计算参数 $\theta$ （频率，样本均值，协方差）： $\ln p(X_L, y_L | \theta) = \sum_{i=1}^L \ln(p(y_i | \theta)p(x_i | y_i, \theta))$

MLE通常计算困难，因此可以用EM算法寻找局部最优解。

核心是最大化 $p(X_L, y_L, X_U | \theta)$

优点：清晰；如果模型接近真实分布会比较有效

缺点：验证模型比较困难；EM局部最优；如果生成式错误的，无监督数据会加重错误。

S<sup>3</sup>VMs，即TSVMs：

基本假设：来自不同类别的无标记数据之间会被较大间隔隔开

基本思想：

- 遍历所有 $2^U$ 种可能性的标注 $X_U$
- 为每一种标注构建一个标准的SVM，包括 $X_L$
- 选择间隔最大的SVM