统计学习复习大纲

Fw[a]rd, SPFA

2026 年 1 月 14 日

1　统计学习问题分类

监督学习：回归、分类；无监督学习：聚类、变换（降维/投影、嵌入）

2　一元统计分析

极大似然　似然函数：$L(\theta)\prod_i p(x^{(i)}|\theta)$，最大化似然函数得参数的极大似然估计 θ_{ML}, 常用办法是对似然函数取负对数，再寻找负对数似然 (NLL) 函数的极小值

贝叶斯法

$$P(\Theta|X^{(1)},...,X^{(n)})=\frac{P(X^{(1)},...,X^{(n)}|\Theta)P(\Theta)}{P(X^{(1)},...,X^{(n)})} \tag{1}$$

先验和后验属于同一类分布的情况在贝叶斯方法中称为共轭先验，是贝叶斯方法中设定先验的一种常见做法

评价准则　统计量：样本的函数

一致性：如果随着样本数量的增大，估计量依概率收敛于真实值，即 $\forall\epsilon>0,\lim_{n\to\infty}P(|\hat{\theta}_n-\theta|<\epsilon)=1$ 则称估计量是（弱）一致的.

无偏性：如果估计量的期望等于真实值，即对任意 θ 有 $\mathbb{E}[\hat{\theta}]=\theta$，则称估计量是无偏的，否则就是有偏的. 将 $\mathbb{E}[\hat{\theta}]-\theta$ 称作估计量的偏差 (bias).

有效性：估计量的方差应该尽可能小. 如果有两个无偏估计量 $\hat{\theta}_1$ 和 $\hat{\theta}_2$，且对任意 θ 有 $\mathbb{V}[\hat{\theta}_1]<\mathbb{V}[\hat{\theta}_2]$，则称 $\hat{\theta}_1$ 比 $\hat{\theta}_2$ 更有效.

充分统计量: 设总体 X 的概率函数带有未知参数 θ，统计量 $\hat{\theta}=f(X^{(1)},...,X^{(n)})$ 其中 f 是预定函数，如果条件概率函数 $\mathbb{P}(X^{(1)},...,X^{(n)}|\hat{\theta})$ 与 θ 无关，则称 $\hat{\theta}$ 是 θ 的充分统计量.

因子分解定理　若样本的概率函数能够分解为 $\mathbb{P}(X^{(1)},...,X^{(n)})=g(x^{(1)},...,x^{(n)})h(f(x^{(1)},...,x^{(n)});\theta)$，则 $\hat{\theta}=f(X^{(1)},...,X^{(n)})$ 是 θ 的充分统计量

指数族分布的充分统计量　若总体服从指数族分布 $P(X)=g(x)h(\theta)\exp\left(\theta^T\phi(x)\right)$, 其中 $\phi(x)$ 称为特征基函数，则 $\sum_{i=1}^n\phi(X^{(i)})$ 是 θ 的充分统计量. 高斯分布 $\phi(x)=(x^2,x)^T$

最小均方误差估计风险分解 $\mathbb{E}[(\hat{\theta}-\theta)^2]=(\mathbb{E}[\hat{\theta}-\theta])^2+\mathbb{V}[\hat{\theta}-\theta]=(\mathbb{E}[\hat{\theta}]-\theta)^2+\mathbb{V}[\hat{\theta}]$ 第一项是估计量的偏差的平方，第二项是估计量的方差．偏差和方差分别反映了估计量的系统误差和随机误差，均方误差最小化同时考虑了系统误差和随机误差．如果将估计量限定为无偏的，则最小均方误差估计量就是一致最小方差无偏估计 (UMVUE)．如果允许估计量有偏，则最小均方误差估计可以不同于 UMVUE.

非参数统计分析 经验分布函数 (EDF):$\hat{F}(x)=\frac{1}{n}\sum_{i=1}^{n}I(x^{(i)}\leq x)$

平滑:$\widetilde{F}(x)\triangleq\int_t\hat{F}(t)k(x-t)\,dt$, k 称为平滑核函数或简称核函数，概率密度函数估计：$\hat{f}(x)=\frac{1}{n}\sum_{i=1}^{n}k\left(x-x^{(i)}\right)$

k 近邻 (k-NN) 法：在实数轴上的每个点 x，调节区间宽度 $h(x)$ 使得区间 $[x-h(x),x+h(x)]$ 中恰好有 k 个数据，则可用 $\hat{f}(x)=\frac{k}{2nh(x)}$ 来估计密度

3 线性回归

最小二乘法

$$\min_{w,b}\mathcal{E}(w,b)=\min_{w,b}\sum_{i=1}^{n}(y^{(i)}-(wx^{(i)}+b))^2 \tag{2}$$

$$w_{LS}=\frac{\sum_{i=1}^{n}(x^{(i)}-\bar{x})(y^{(i)}-\bar{y})}{\sum_{i=1}^{n}(x^{(i)}-\bar{x})^2},b_{LS}=\bar{y}-w_{LS}\bar{x} \tag{3}$$

正则化 带约束优化问题：

$$\min_{w}\sum(y^{(i)}-wx^{(i)})^2,s.t.w^2\leq c \tag{4}$$

$$L(w,\lambda)=\sum(y^{(i)}-wx^{(i)})^2+\lambda(w^2-c) \tag{5}$$

$$w_{reg}=\frac{\sum x^{(i)}y^{(i)}}{\sum(x^{(i)})^2+\lambda} \tag{6}$$

正则化是在求最小二乘解的时候限制参数 w 的取值范围，正则化权重越大，取值范围限定得越小．

偏差-方差均衡:

$$\mathbb{E}[\widehat{w}]=\frac{\sum(x^{(i)})^2}{\sum(x^{(i)})^2+\epsilon}w,\mathbb{V}[\widehat{w}]=\frac{\sum(x^{(i)})^2\sigma^2}{(\sum(x^{(i)})^2+\epsilon)^2} \tag{7}$$

ϵ 越大，$\mathbb{E}[\widehat{w}]$ 偏离 w 越多，偏差平方越大；$\mathbb{V}[\widehat{w}]$ 越小，方差越小．综合考虑偏差和方差，则可以找到一个合适的 ϵ，使得两项之和达到最小

贝叶斯：$W\sim N(0,\sigma_w^2)$,$P(Y^{(i)}|W)\sim N(wx^{(i)},\sigma^2)$, 则 $\epsilon=\sigma^2/\sigma_w^2$，正则化项实质上对应于后验分布中由先验分布引入的项

最小二乘回归

$$X=\begin{bmatrix}(x^{(1)})^T\\ \vdots\\ (x^{(n)})^T\end{bmatrix} \tag{8}$$

$$w_{LS}=arg\min_{w}(y-Xw)^T(y-Xw) \tag{9}$$

$$X^Ty=X^TXw \tag{10}$$

带基函数的回归

$$\Phi \triangleq (f_1, ..., f_p)^T, y = w^T \Phi(x) \tag{11}$$

$$\Phi^T \Phi w_{LS} = \Phi^T y \tag{12}$$

岭回归

$$\min_w (y - Xw)^T (y - Xw) + \lambda w^T w \tag{13}$$

$$w_{ridge} = (X^T X + \lambda I)^{-1} X^T y \tag{14}$$

贝叶斯线性回归 假设 $P(W) = N(m_0, S_0)$，样本条件分布如下，可得

$$P(Y^{(1)}, ..., Y^{(n)}|W) = \prod N(w^T x^{(i)}, \sigma^2) \tag{15}$$

$$P(W|Y^{(1)}, ..., Y^{(n)}) = N(m_n, S_n) \tag{16}$$

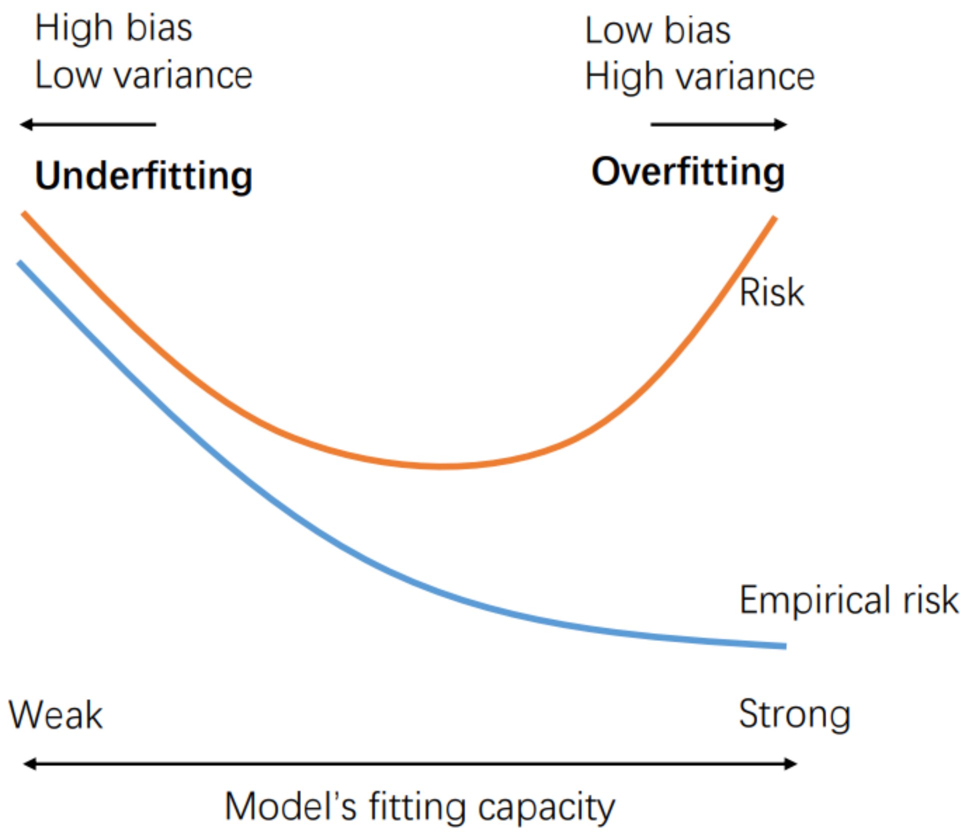
$$m_n = S_n(S_0^{-1} m_0 + \frac{1}{\sigma^2} X^T y) \tag{17}$$

$$S_n = (S_0^{-1} + \frac{1}{\sigma^2} X^T X) \tag{18}$$

序贯学习：

$$\begin{gathered} P(Y^{(1)}, ..., Y^{(n+1)}|W) = \\ \frac{P(Y^{(1)}, ..., Y^{(n)}|W) P(Y^{(n+1)}|W)}{P(Y^{(n+1)}|Y^{(1)}, ..., Y^{(n)})} \end{gathered} \tag{19}$$

模型评价与选择 一般来说，如果模型中可学习参数太少，模型的拟合能力很弱，在训练数据上的经验风险很大，风险也很大，这种现象称为欠拟合. 经验风险降低但风险反而升高的现象在统计学习中称为过拟合



赤池信息量准则 $AIC = 2NLL + 2p$, 其中，NLL 是训练数据上估计的负对数似然，是参数的个数．AIC 值越小，模型越好．当 p 一定时，负对数似然越小的回归函数越好，也就是说极大似然估计是最好的，代入得 $AIC = nln(\mathcal{E}(w_{LS})) + 2p$

贝叶斯模型评价 对参数（随机变量）求期望，称为模型证据

LASSO

$$\min_w (y - Xw)^T(y - Xw) + \lambda\|w\|_1 \tag{20}$$

lasso 回归虽然具有特征选择等优点，但它一般情况下没有闭式解

核回归

$$\widehat{f}(x) \triangleq \sum_{i=1}^{n} y^{(i)} k(x, x^{(i)}) \tag{21}$$

Mercer 条件：令矩阵 K，其中 $K_{ij} = k(x^{(i)}, x^{(j)})$，要求对任意 x, K 半正定．有 $k(x, y) = k(y, x)$，$k(x, y) = (\Phi(x))^T\Phi(y)$

$$\widehat{f}(x) \triangleq \sum_{i=1}^{n} y^{(i)} \Phi(x^{(i)})^T \Phi(x) = y^T \Phi\Phi(x) = w^T \Phi(x) \tag{22}$$

K 近邻回归

$$\widehat{f}(x) \triangleq \frac{1}{k}\sum_{j=1}^{k} y^{(n_j)}, s.t.x^{(n_j)} \in N_k(x) \tag{23}$$

4 线性分类

线性分类

$$y = \operatorname{sign}(w_{LS}^T x + b_{LS}) \tag{24}$$

有 train-test mismatch 问题

zero-one loss/0-1 loss 考虑到 sign 函数，平方损失函数等价于

$$L(w,b,x,y) \triangleq I(y \neq sign(w^T x + b)) \tag{25}$$

Fisher 投影 (LDA) 对于 +1 类，v_+ 为均值，$m_+ = w^T v_+$ 为投影后均值，$\boldsymbol{S}_+$ 为协方差矩阵，$S_+ = w^T \boldsymbol{S}_+ w$ 为投影后方差. -1 类同理. 类内方差 $S_w = S_+ + S_-$ 类间方差 $(m_+ - m_-)^2$，求：

$$\min_w w^\mathsf{T}(\boldsymbol{S}_+ + \boldsymbol{S}_-)w, \text{s.t.} w^\mathsf{T}(v_+ - v_-)(v_+ - v_-)^\mathsf{T} w \geqslant C \tag{26}$$

$$w \propto (\boldsymbol{S}_+ + \boldsymbol{S}_-)^{-1}(v_+ - v_-) \tag{27}$$

感知机 Perceptron 使用代理损失函数：

$$L(w,b,x,y) \triangleq \max(0, -y(w^T x + b)) \tag{28}$$

使用随机梯度下降 SGD

$$if y_i(w^T x_i + b) < 0 : w = w + \eta y_i x_i, b = b + \eta y_i \tag{29}$$

对偶形式：$w = \sum \alpha_i y_i x_i, b = \sum \alpha_i y_i, \alpha_i \geq 0$，直接学 α_i

$$if y_i(\sum \alpha_j y_j x_j^T x_i + \sum \alpha_j y_j) < 0 : \alpha_i = \alpha_i + \eta \tag{30}$$

最终分类器为 $sign(\sum \alpha_i y_i (x_i^T x + 1))$

带基函数感知机

$$if y_i((\sum \alpha_j y_j \Phi(x_j))^T \Phi(x_i)) < 0 : \alpha_i = \alpha_i + \eta \tag{31}$$

最终分类器为 $sign((\sum \alpha_i y_i \Phi(x_i))^T \Phi(x))$

带核函数感知机

$$if \sum_j \alpha_j y_i y_j k(x_i, x_j) < 0 : \alpha_i = \alpha_i + \eta \tag{32}$$

最终分类器为 $sign(\sum \alpha_i y_i k(x_i, x))$

交叉熵损失

$$CE(B(p), B(q)) = -plnq - (1-p)ln(1-q) \tag{33}$$

逻辑回归 样本 (x_i, y_i),$y_i \sim B(q_i)$,$q_i = f(x_i; w, b)$，令

$$q_i = \frac{1}{1 + exp(-w^T x_i - b)} \triangleq \sigma(w^T x_i + b) \tag{34}$$

$\sigma(x) = \frac{1}{1+e^-x}$+ 交叉熵损失 = 逻辑回归

广义逻辑回归 GLR

$$\min_w \sum(-y_i lnq_i - (1-y_i)ln(1-q_i)) \tag{35}$$

其中 $q_i = \frac{1}{1+exp(-w^T \Phi(x_i))}$

分类器为 $y = sign(-w^T \Phi(x))$

GLR 的解

$$L(w) = \sum(-y_i lnq_i - (1-y_i)ln(1-q_i)) \tag{36}$$

$$\nabla L(w) = \sum(q_i - y_i)\mathit{\Phi}(x_i) = \Phi^T(q-y) \tag{37}$$

$$\nabla\nabla L(w) = \sum q_i(1-q_i)\mathit{\Phi}(x_i)\mathit{\Phi}^T(x_i) = \Phi^T R \Phi \tag{38}$$

其中 $R = diag(q_i(1-q_i))$，根据牛顿迭代法

$$w^{new} = w^{old} - (\Phi^T R \Phi)^{-1}\Phi^T(q-y) \tag{39}$$

令 $z \triangleq \Phi w^{old} - R^{-1}(q-y)$，有

$$w^{new} = (\Phi^T R \Phi)^{-1}\Phi^T R z \tag{40}$$

w^{new} 可视为加权最小二乘问题的解

$$\min_w (z - \Phi w)^T R (z - \Phi w) \tag{41}$$

当 q_i 远离 0.5 时权重变低

朴素贝叶斯

$$P(Y=i|X) = \frac{P(Y=i)P(X|Y=i)}{\sum_{i=1}^n P(Y=i)P(X|Y=i)} \tag{42}$$

其中 $P(X|Y=i) = \prod_j P(X_j|Y=i)$

K-NN

$$\widehat{f}(x) \triangleq sign(\frac{1}{k}\sum_{j=1}^{k} y^{(n_j)}), s.t. x^{(n_j)} \in N_k(x) \tag{43}$$

稀疏表示

$$x \approx \sum_{i=1}^{n} \alpha_i x^{(i)}, s.t. \sum_{i=1}^{n} I(\alpha_i \neq 0) = k \tag{44}$$

$\alpha = (\alpha_1, ..., \alpha_n)^T$ 称为 K-稀疏向量，分类器为

$$\widehat{f}(x) \triangleq sign(\frac{1}{k}\sum_{j=1}^{k} y^{(n_j)}), s.t. \alpha_{n_j} \neq 0 \tag{45}$$

解 α:

$$\min_{\alpha} (x - X\alpha)^T (x - X\alpha) + \lambda \|\alpha\|_1 \tag{46}$$

其中 $X = (x^{(1)}, ..., x^{(n)})$，把 2-范数换成 1-范数也行

硬边界 SVM 假设数据线性可分，分类器无误差，分类边界 $w^T x_i + b = 0$，则 $y_i = sign(w^T x_i + b)$，距离可写为

$$d_i = \frac{|w^T x_i + b|}{\|w\|_2} = \frac{y_i(w^T x_i + b)}{\|w\|_2} \tag{47}$$

最大间隔问题为 $\max_{w,b} \min_i d_i$，记为

$$\max_{w,b} \gamma(w, b), \gamma(w, b) \triangleq \min_i \frac{y_i(w^T x_i + b)}{\|w\|_2} \tag{48}$$

考虑到 $\forall \alpha > 0, \gamma(\alpha w, \alpha b) = \gamma(w, b)$，问题化为

$$\max_{w,b} \frac{1}{\|w\|_2}, min_i y_i(w^T x_i + b) = 1 \tag{49}$$

进一步松弛为

$$\min_{w,b} \frac{1}{2}\|w\|_2^2, s.t. \forall i, y_i(w^T x_i + b) \geq 1 \tag{50}$$

拉格朗日乘子 $\alpha = (\alpha_1, ..., \alpha_n)^T$，拉格朗日函数为

$$L(w, b, \alpha) = \frac{1}{2}\|w\|_2^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i(w^T x_i + b)) \tag{51}$$

KKT 条件为

$$\nabla_w L(w, b, \alpha) = 0 \Rightarrow w = \sum_{i=1}^{n} \alpha_i y_i x_i \tag{52}$$

$$\frac{\partial}{\partial b} L(w, b, \alpha) = 0 \Rightarrow \sum_{i=1}^{n} \alpha_i y_i = 0 \tag{53}$$

$$\alpha_i \geq 0, i = 1, ..., n \tag{54}$$

$$1 - y_i(w^T x_i + b) \leq 0, i = 1, ..., n \tag{55}$$

$$\alpha_i (1 - y_i(w^T x_i + b)) = 0, i = 1, ..., n \tag{56}$$

当 $\alpha_i > 0$ 时，距离满足 $y_i(w^T x_i + b) = 1$，只有这些点对计算 w 有贡献，称为支持向量

软边界 SVM 数据通常不线性可分；有时可分，但为扩大间隔，去掉一些.

$$\min_{w,b}\frac{1}{2}\|w\|_2^2, s.t. \sum_{i=1}^{n} I(y_i(w^T x_i + b) < 1) \leq c \tag{57}$$

用代理损失函数 $\max(1 - y_i(w^T x_i + b), 0)$，用拉格朗日法

$$\min_{w,b}\frac{1}{2}\|w\|_2^2 + \lambda\sum_{i=1}^{n}\max(1 - y_i(w^T x_i + b), 0) \tag{58}$$

令 $\xi_i = \max(0, 1 - y_i(w^T x_i + b))$，再松弛为

$$\begin{gathered}\min_{w,b}\frac{1}{2}\|w\|_2^2 + \lambda\sum_{i=1}^{n}\xi_i, s.t. \forall i, \xi_i \geq 1 - y_i(w^T x_i + b),\\ \xi_i \geq 0\end{gathered} \tag{59}$$

拉格朗日函数为

$$\begin{aligned}L(w,b,\xi,\alpha,\beta) = \frac{1}{2}\|w\|_2^2 + \lambda\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\beta_i\xi_i\\ - \sum_{i=1}^{n}\alpha_i(y_i(w^T x_i + b) - 1 + \xi_i)\end{aligned} \tag{60}$$

KKT 条件为

$$\nabla_w L(w,b,\xi,\alpha,\beta) = 0 \Rightarrow w = \sum_{i=1}^{n}\alpha_i y_i x_i \alpha_i y_i = 0 \tag{61}$$

$$\frac{\partial}{\partial b}L(w,b,\xi,\alpha,\beta) = 0 \Rightarrow \sum_{i=1}^{n} \tag{62}$$

$$\frac{\partial}{\partial \xi}L(w,b,\xi,\alpha,\beta) = 0 \Rightarrow \alpha_i + \beta_i = \lambda \tag{63}$$

$$\alpha_i \geq 0, \beta_i \geq 0, i = 1, ..., n \tag{64}$$

$$\xi_i \geq 1 - y_i(w^T x_i + b), \xi_i \geq 0, i = 1, ..., n \tag{65}$$

$$\alpha_i(y_i(w^T x_i + b) - 1 + \xi_i) = 0, i = 1, ..., n \tag{66}$$

$$\beta_i\xi_i = 0, i = 1, ..., n \tag{67}$$

代掉 β_i，得

$$\begin{aligned}&w = \sum_{i=1}^{n}\alpha_i y_i x_i\\ &\sum_{i=1}^{n}\alpha_i y_i = 0\\ &0 \leq \alpha_i \leq \lambda, i = 1, ..., n\\ &\xi_i \geq 1 - y_i(w^T x_i + b), \xi_i \geq 0, i = 1, ..., n\\ &\alpha_i(y_i(w^T x_i + b) - 1 + \xi_i) = 0, i = 1, ..., n\\ &(\lambda - \alpha_i)\xi_i = 0, i = 1, ..., n\end{aligned} \tag{68}$$

所有距离可分为 3 类

$$
\begin{aligned}
& y_i(w^Tx_i+b)>1 \Rightarrow \alpha_i=0, \xi_i=0 \\
& y_i(w^Tx_i+b)=1 \Rightarrow 0<\alpha_i<\lambda, \xi_i=0 \\
& y_i(w^Tx_i+b)<1 \Rightarrow \alpha_i=\lambda, \xi_i>0
\end{aligned} \tag{69}
$$

当 $\alpha_i>0$ 时，距离满足 $y_i(w^Tx_i+b)\leq 1$，只有这些点对计算 w 有贡献，称为支持向量

对偶问题 $\max_\alpha \min_{w,b,\xi} L(w,b,\xi,\alpha)$，由 KKT 条件得

$$
\begin{aligned}
\max_\alpha -\frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_j y_i y_j x_i^T x_j + \sum_{i=1}^n \alpha_i \\
\text{s.t. } 0\leq \alpha_i \leq \lambda, \sum_{i=1}^n \alpha_i y_i = 0
\end{aligned} \tag{70}
$$

分类器为 $y=sign(\sum_{i=1}^n \alpha_i y_i x_i^T x+b)$，$b$ 由 $0<\alpha_i<\lambda$ 的支持向量计算得出

带基函数 SVM & 核 SVM

$$
\begin{aligned}
\min_{w,b} \frac{1}{2}\|w\|_2^2 + \lambda\sum_{i=1}^n \xi_i, s.t. \forall i, \xi_i \geq 0, \\
\xi_i \geq 1-y_i(w^T\Phi(x_i)+b)
\end{aligned} \tag{71}
$$

KKT 条件为：

$$
\begin{aligned}
& w=\sum_{i=1}^n \alpha_i y_i \Phi(x_i) \\
& \sum_{i=1}^n \alpha_i y_i = 0 \\
& 0\leq \alpha_i \leq \lambda, i=1,...,n \\
& \xi_i \geq 1-y_i(w^T\Phi(x_i)+b), \xi_i\geq 0, i=1,...,n \\
& \alpha_i(y_i(w^T\Phi(x_i)+b)-1+\xi_i)=0, i=1,...,n \\
& (\lambda-\alpha_i)\xi_i=0, i=1,...,n
\end{aligned} \tag{72}
$$

对偶问题为

$$
\begin{aligned}
\max_\alpha -\frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_j y_i y_j \Phi(x_i)^T\Phi(x_j) + \sum_{i=1}^n \alpha_i \\
\text{s.t. } 0\leq \alpha_i \leq \lambda, \sum_{i=1}^n \alpha_i y_i = 0
\end{aligned} \tag{73}
$$

分类器为 $y=sign(\sum_{i=1}^n \alpha_i y_i \Phi(x_i)^T\Phi(x)+b)$

核函数 $k(x_i,x_j)=\Phi(x_i)^T\Phi(x_j)$

序列最小优化算法 SMO 随机选择两个 α_p 和 α_q，固定其他 α_i，在(73)中 $arg\max_{\alpha_p,\alpha_q}$

比较 贝叶斯分类规则: 对于 (X,Y), 使用 0-1loss, 最佳分类器为

$$f^*(x) = arg\min_f E_{X,Y}[I(Y \neq f(X))] = \\ arg\max_{k\in\{+1,-1\}} P(Y=k|X=x) \tag{74}$$

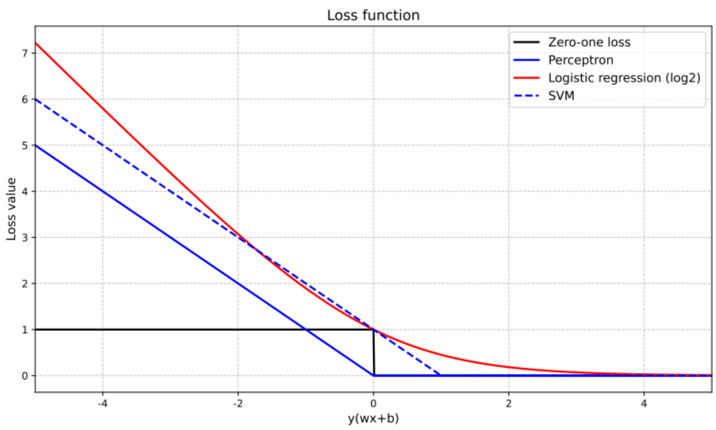
LDA 假设各类服从高斯分布，协方差矩阵相同

GLR 假设各类服从指数族分布

朴素贝叶斯认为 $P(X|Y)=\prod_j P(X_j|Y)$

K-NN 认为贝叶斯最优分类器在局部是常数

方法	损失函数
感知机	$\max(-y(w^Tx+b),0)$
逻辑回归	$ln(1+exp(-y(w^Tx+b)))$
SVM	$\max(1-y(w^Tx+b),0)$



对 0-1loss 的上界进行优化

$0-r_{+1}-r_{-1}loss$ +1 类误分损失 r_{+1}，-1 类误分损失 r_{-1}

Precision：$\frac{TP}{TP+FP}$；Recall/TPR：$\frac{TP}{TP+FN}$；FPR：$\frac{FP}{FP+TN}$；F1-value：$2\cdot\frac{Precision\cdot Recall}{Precision+Recall}$；ROC 曲线：以 FPR 为横轴，TPR 为纵轴绘制的曲线；AUC：ROC 曲线下的面积

正则化 考虑到几何解释，SVM 本身就使用 L2 正则化；逻辑回归自然可以使用 L1 或 L2 正则化；对于朴素贝叶斯，可以使用拉普拉斯平滑

基函数和核函数 所有方法均可用基函数 (视作预处理);SVM 和感知机可用核函数;K-NN 可用核函数定义距离

Softmax 回归 设 $P(Y=k)=1/K, P(X|Y=k)=exp(\theta_k^T\Phi(x))$ 则

$$P(Y=k|X=x)=\frac{exp(\theta_k^T\Phi(x))}{\sum_l exp(\theta_l^T\Phi(x))} \tag{75}$$

问题形式为

$$\min_{\theta} - \sum_{i=1}^{n}\sum_{k=1}^{K} I(y_i = k) ln \frac{exp(\theta_k^T \Phi(x))}{\sum_l exp(\theta_l^T \Phi(x))} \tag{76}$$

分类器为 $y = \arg\max_k \theta_k^T \Phi(x)$

带温度 Softmax

$$softmax(z_k, \tau) = \frac{exp(z_k/\tau)}{\sum_l exp(z_l/\tau)} \tag{77}$$

sigmoid 为 sign 的 soft 版本，softmax 为 $I(z_k = max_k z_k)$ 的 soft 版本．τ 越小，softmax 越接近后者．模拟退火先用大 τ，再慢慢减小

5 无监督学习

维度诅咒 高维空间中难以估计 (概率) 密度: 样本数不够; 邻居太远; 距离难分辨．若样本数为 n^p 则无问题，但难以承担

降维/投影 $x_i \in \mathbb{R}^p$ 寻找映射 $f : \mathbb{R}^p \to \mathbb{R}^k$，$p < k$

解决维度灾难; 减少数据量及计算成本; 去噪声及无关特征, 避免过拟合

PCA 先中心化，再用平方损失寻找投影 $f(x) = Ex, E \in \mathbb{R}^{p\times k}$ 反投影 $g(z) = Dz, D \in \mathbb{R}^{k\times p}$

$$(E^*, D^*) = \arg\min_{E,D} \sum_{i=1}^{n} \|x_i - DEx_i\|_2^2 \tag{78}$$

对于 E^T 和 D 的列向量 e_j, d_j, 令 $z_{ij} = e_j^T x_i$ 化为

$$\sum_{i=1}^{n} \left\| x_i - \sum_{j=1}^{k} z_{ij} d_j \right\|_2^2 \tag{79}$$

假设 d_j 正交, 得到 $z_{ij} = d_j^T x_i$, 即 $e_j = d_j$, 原问题化为

$$D^* = \arg\min_{D} \sum_{i=1}^{n} \left\| x_i - DD^T x_i \right\|_2^2, s.t. D^T D = I \tag{80}$$

令 $L = \sum_{i=1}^{n} \left\| x_i - DD^T x_i \right\|_2^2$, 得

$$\begin{aligned} L = tr(X^T X) - tr(D^T X^T X D) = \\ tr(X^T X) - \sum_{j=1}^{k} d_j^T X^T X d_j \end{aligned} \tag{81}$$

原问题用拉格朗日法得

$$D^* = \arg\min_{D} - \sum_{j=1}^{k} d_j^T X^T X d_j + \sum_{j=1}^{k} \lambda_j (d_j^T d_j - 1) \tag{82}$$

$$\forall j, -2X^T X d_j + s\lambda_j d_j = 0 \Rightarrow X^T X d_j = \lambda_j d_j \tag{83}$$

D 为 $X^\mathsf{T} X = \sum x_i x_i^\mathsf{T}$ 的前 k 个标准化特征向量, 投影 $z = D^\mathsf{T} x = (d_1^\mathsf{T} x, ..., d_k^\mathsf{T} x)^\mathsf{T}$, 反投影 $\widehat{x} = Dz + \overline{x} - DD^\mathsf{T} \overline{x}$

几何意义: 旋转 (去相关) 后取方差最大的几个分量．尽可能保持最大方差

核 PCA 用 $\mathit{\Phi}(x)$ 代替 x，相当于计算" 相似性"

多维尺度分析 MDS 定义距离 $d_{ij} = dist(x_i, x_j)$ 问题为

$$\min_{z_1,\ldots,z_n} \sum_i \sum_j (\|z_i - z_j\| - d_{ij})^2 \tag{84}$$

流形 全局非线性, 局部线性. 若两点为邻居, 用欧几里得距离. 若不是, 用测地距离 (连接两点且位于流形上的最短线段的长度)

ISOMAP 建立一个图, 顶点为数据, 邻居间连边, 边权为欧几里得距离. 测地距离通过最短路算法计算, 最后用 MDS

局部线性嵌入 LLE 定义 $W \in \mathbb{R}^{n\times n}$, 若 x_i, x_j 不是邻居则 $W_{i,j} = 0$, 优化 W 的其它元素

$$\min_W \sum_{i=1}^n \left\| x_i - \sum_j W_{i,j} x_j \right\|^2 s.t. \sum_j W_{i,j} = 1 \tag{85}$$

再计算投影

$$\min_{z_1,\ldots,z_n} \sum_{i=1}^n \left\| z_i - \sum_j W_{i,j} z_j \right\|^2 \tag{86}$$

K-MEANS 令 $\mathcal{K} = 1, .., k$, 聚类要找 $f : \mathbb{R}^p \to \mathcal{K}$, 其反函数 $g : \mathcal{K} \to \mathbb{R}^p.g$ 可用 k 个常向量 $c_1, ..., c_k$ 表示, 称为 codebook, 每个向量称为 codeword. 用平方损失函数, 经验风险为

$$\mathcal{E}(f, c_1, ..., c_k) = \sum_{i=1}^n \left\| x_i - c_{f(x_i)} \right\|_2^2 \tag{87}$$

K-MEANS 用启发式迭代算法优化 $\mathcal{E}$, 到收敛为止

1. 若 c 已知,$f(x_i) = \arg\min_j \|x_i - c_j\|_2^2$
2. 若 f 已知,$c_j = \frac{\sum_i I(f(x_i)=j)x_i}{\sum_i I(f(x_i)=j)}$

通常用多个不同的初始化训练，取最好的. 可先过分类 (用更多的 c) 再后处理提升表现

高斯混合模型 GMM 第 j 簇服从高斯分布 $N(\mu_j, \Sigma_j)$. 令 X_i 为样本, 对应簇为 Z_i,Z_i 间 i.i.d.

$$P(Z_i) = \prod_{j=1}^k w_j^{I(z_i=j)}, \sum_{j=1}^k w_j = 1 \tag{88}$$

$$P(X_i|Z_i) = \prod_{j=1}^k (N(X_i; \mu_j, \Sigma_j))^{I(z_i=j)} \tag{89}$$

$$P(X_i) = \sum_{j=1}^k w_j N(X_i; \mu_j, \Sigma_j) \tag{90}$$

X_ii.i.d.

期望最大化算法 (EM) for GMM 已知参数 w_j, μ_j, Σ_j

$$P(Z_i = j|X_i) = \frac{w_j N(x_i; \mu_j, \Sigma_j)}{\sum_{j=1}^{k} w_j N(x_i; \mu_j, \Sigma_j)} \triangleq \gamma_{ij} \tag{91}$$

$$\begin{aligned} w_j &= \frac{\sum_i \gamma_{ij}}{n}, \mu_j = \frac{\sum_i \gamma_{ij} x_i}{\sum_i \gamma_{ij}} \\ \Sigma_j &= \frac{\sum_i \gamma_{ij}(x_i - \mu_j)(x_i - \mu_j)^\top}{\sum_i \gamma_{ij}} \end{aligned} \tag{92}$$

迭代(91)和(92)直到收敛

$$f(x_i) = \arg\max_j P(Z_i = j|X_i = x_i) = \arg\max_j \gamma_{ij} \tag{93}$$

K-MEANS 与 GMM K-MEANS 是 GMM 的特例，认为 $w_j = 1/k, \Sigma_j = I$,γ_{ij} 计算如下

$$\begin{aligned} \gamma_{i,j} =& 1, if P(Z_i = j|X_i) = \max_l P(Z_i = l|X_i) \\ & 0, otherwise \end{aligned} \tag{94}$$

k-means 使用硬分配 (hard assignment),GMM-EM 用软分配, 因此 k-means 对于具有不同大小、密度或不规则形状的簇存在局限性

EM 解决带隐变量的最大似然估计问题．观测变量 X, 隐变量 Z, 待估参数 θ

算法 1 EM 算法

1: $t \leftarrow 0$, initialize θ^0
2: **repeat**
3: E-step: $Q(\theta) = \mathbb{E}_{Z\sim P(Z|X=x,\theta^t)}[\log P(X, Z; \theta)]$
4: M-step: $\theta^{t+1} = \arg\max_\theta Q(\theta)$
5: **until** $\|\theta^{t+1} - \theta^t\| < \epsilon$
6: $\hat{\theta} = \theta^{t+1}$

$$\begin{aligned} logP(X;\theta) &= (\sum_z P(Z = z|X;\theta^t))logP(X;\theta) \\ &= \sum_z P(Z = z|X;\theta^t)logP(X, Z;\theta) \\ &- \sum_z P(Z = z|X;\theta^t)logP(Z|X;\theta) \\ &\triangleq Q(\theta) + H(\theta) \end{aligned} \tag{95}$$

$H(\theta)$ 是 $P(Z|X;\theta^t)$ 和 $P(Z|X;\theta)$ 间的交叉熵, 有 $H(\theta) \geq H(\theta^t)$. 我们优化 $Q(\theta)$, 有 $Q(\theta^{t+1}) \geq Q(\theta^t)$. 因此有 $logP(X;\theta^{t+1}) \geq logP(X;\theta^t)$.EM 为贪心, 每一步 P 不减

非参数聚类

基于距离的聚类 凝聚聚类: 自底向上, 合并相近数据或簇; 分离聚类: 自顶向下, 通过切掉距离最长的边来分割子图

基于距离的聚类 Mean-shift: 用 Parzen 窗估计局部密度并计算局部均值, 将局部模式 (密度最高的点) 移到均值处;DBSCAN: 给定一个随机选择的数据，找到它的最近邻居并估计局部密度；如果密度足够高，则将此数据及邻居设置为簇，并尝试扩展簇，直到到达低密度区域

嵌入 embedding 增加特征的维度，或为对象构建高维特征向量

例: 评分预测, 设有 n 个电影和 m 个用户, 每部电影有 1 个嵌入向量 $m_i \in \mathbb{R}^p$, 每个用户有 1 个嵌入向量 $u_i \in \mathbb{R}^p$, 评分为 $r_{ij} = m_i^{\mathbf{T}} u_j$, 评分矩阵为 $R = M^{\mathbf{T}} U$, 为一个低秩矩阵. 若已知 R, 可用截断 SVD 得到 M, U 用于评分预测

例: 词嵌入, 可用 LDA 思路 (减小类内方差, 增大类间差别)

6 基于树的模型与集成学习

回归树桩 regression stump

$$f(x) = (\beta_1 - \beta_0) sign(w^T x + b) + \beta_0 \tag{96}$$

模型组合 线性组合线性模型等价于另一线性模型, 一般不如此组合. 若基模型表现较好且有多样性 (well and diversely), 则组合模型一定有提升.

保证 well and diversely 的方法: 训练数据多样性 (数据分割、特征分割、不同核); 训练方法多样性

性能和多样性存在矛盾

模型组合方法: 简单相加/投票 (bagging,boosting); 训练组合 (stacking: 每个基模型学一个特征, 再训练一个总的模型进行组合); 局部自适应组合 (树模型: 将输入空间分为若干子域, 每个基模型处理一个)

bootstrap aggregating(bagging) 使用 bootstrap sampling(自助抽样) 得到数据多样性: 给定数据集 $x^{(i)}|_{i=1}^n$, 进行 n 次带放回的均匀采样.

一个数据抽不到的概率为 $(1 - \frac{1}{n})^n$, 当 $n \to +\infty$, 有 $1/e$ 的数据抽不到. 使用一次自助抽样训练一个基模型再组合起来.

Boosting 多个基模型一个一个地训练, 组合起来的模型会一点一点变好.

Boosting for 回归

$$F_p(x) = \sum_{j=1}^{p} \beta_j f_j(x) \tag{97}$$

其中 f_j 为基模型, F_p 为总的模型.Boosting 中模型一个一个训练, 可考虑

$$F_j(x) = F_{j-1}(x) + f_j(x) \tag{98}$$

使用平方损失函数

$$\begin{aligned} L(f_j) &= \sum_{i=1}^{n}(y_i - F_j(x_i))^2 \\ &= \sum_{i=1}^{n}(y_i - F_{j-1}(x_i) - f_j(x_i))^2 \end{aligned} \tag{99}$$

令 $r_i = y_i - F_{j-1}(x_i)$, 则 f_j 在回归 $(x^{(i)}, r^{(i)})$, 即除 f_1 外,f_i 在回归残差

算法 2 AdaBoost Algorithm

Require: $(x_i, y_i)|_{i=1}^{n}$,$y_i \in +1, -1$
Ensure: $F_p(x) = sign(\sum_{j=1}^{p} \beta_j f_j(x))$
for $j = 1, \ldots, p$ **do**
  if j=1 **then**
    $w_{ij} = 1/n$
  else
    $w_{ij} = w_{i,j-1} exp(-y_i \beta_{j-1} f_{j-1}(x_i))$
    $w_{ij} = \frac{w_{ij}}{\sum_i w_{ij}}$
  用 w_{ij} 给第 i 个数据加权, 训练分类器 f_j
  计算加权错误率 $e_j = \sum_i w_{ij} I(y_i \neq f_j(x_i))$
  $\beta_j = \frac{1}{2} ln \frac{1-e_j}{e_j}$

AdaBoost 在加权数据上训练的两种方法

1. 对 loss 加权, 如 $\sum_{i=1}^{n} w_{ij}(-y_i lnq_i - (1 - y_i) ln(1 - q_i))$
2. 无显式 loss, 可将 w_{ij} 作为概率, 对数据重采样

若 f_j 对 x_i 分类正确, 则权重下降 $(\times exp(-\beta_j))$, 反之上升 $(\times exp(\beta_j)).f_{j+1}$ 将更关注分类错误的数据.

实际上 AdaBoost 使用指数损失函数, 设分类器 $y = sign(x) \in +1, -1$

$$L(f; x, y) = exp(-yf(x)) \tag{100}$$

最小化 e_j 实际上就是在加权数据上训练一个基分类器, 指数损失函数也是 0-1loss 的一个上界

决策树 一棵树, 每个内部节点对应某些特征的条件, 每个叶节点表示一类 (分类) 或一个值 (回归)

算法 3 HA

Require: A set of training data $\mathcal{D}=\{x_i,y_i\}$

Ensure: A classification tree or regression tree T

```
function HA(D)
    if D 不用分裂 then
        return 叶节点
    else
        寻找一个条件来分裂
        将 D 按条件分裂为 D_1, D_2, ...,
        T_1 = HA(D_1), T_2 = HA(D_2), ...
        建树, 条件为根, T_1, T_2, ... 为子树
    return 生成的树
```

Hunt's algotithm(HA)

分裂条件选取 贪心: 最小化当前经验风险

回归树: 若使用平方损失

$$\min_{j,t_j,\alpha_j,\beta_j}\sum_{i=1}^{n}(\alpha_j I(x_j^{(i)}<t_j)+\beta_j I(x_j^{(i)}>t_j)-y^{(i)})^2 \tag{101}$$

取 α_j 为 $\{y^{(i)}|x_j^{(i)}<t_j\}$ 的均值,β_j 同理

二元分类树: 设 p_0 为 0 的百分数,p_1 同理, 常用 3 个指标:

误分率:$\mathcal{E}(D)=min(p_0,p_1)$

熵:$H(D)=-p_1log_2p_1-p_0log_2p_0$

Gini 指数:$G(D)=1-p_1^2-p_0^2$

使用获得信息量决定分裂几支, 定义如下:

$$r\triangleq\frac{H(D)-\sum_i\frac{|D_i|}{|D|}H(D_i)}{-\sum_i\frac{|D_i|}{|D|}log_2\frac{|D_i|}{|D|}} \tag{102}$$

防止过拟合 早停: 提前停止分裂, 即使还能分

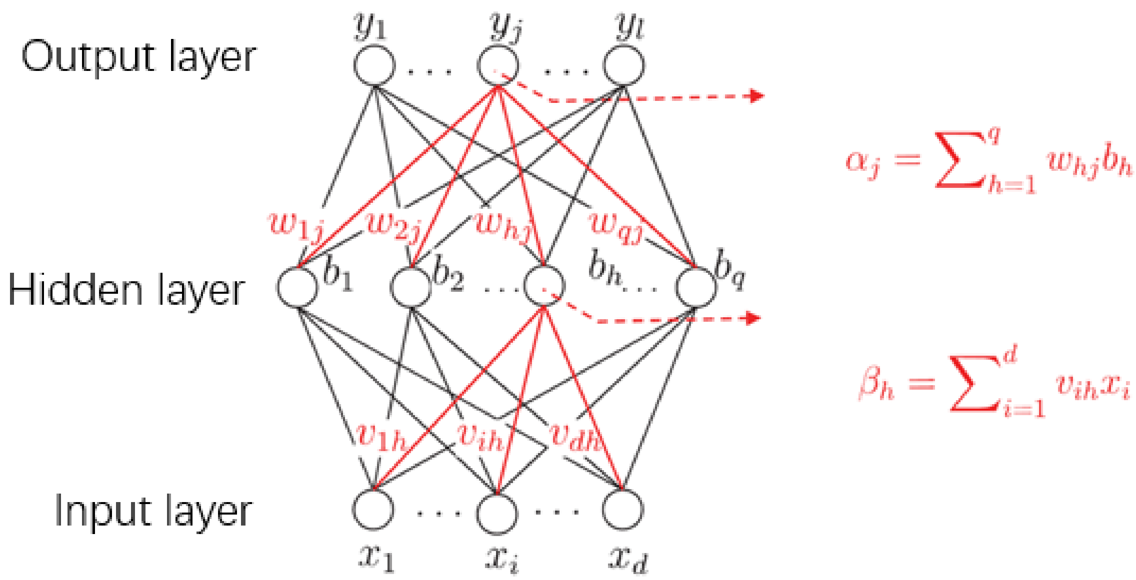
剪枝: 从训练好的树上移除树枝

一般来说剪枝优于早停, 剪枝考虑联合成本:

$$J(T)\triangleq\mathcal{E}(D,T)+\lambda|T| \tag{103}$$

$\mathcal{E}$ 表示经验风险,$|T|$ 表示树的复杂度

7 图模型和深度学习



BP 网络模型 训练数据 $D = \{(x_k, y_k)\}, x_k \in \mathbb{R}^d, y_k \in \mathbb{R}^l$
待学习参数: 权重:v_{ih}, w_{hj}; 偏置:γ_h, θ_j

梯度下降 GD 给定样本 (x^k, y^k), 输出 $\hat{y}^k$

$$
\begin{aligned}
b_h &= f(\beta_h - \gamma_h), \beta_h = \sum_{i=1}^{d} v_{ih} x_i^k \\
\hat{y}_j^k &= f(\alpha_j - \theta_j), \alpha_j = \sum_{h=1}^{q} w_{hj} b_h \\
E_k &= \frac{1}{2} \sum_{j=1}^{l} (\hat{y}_j^k - y_j^k)^2
\end{aligned} \tag{104}
$$

每个参数 ν 更新为

$$
\begin{aligned}
\nu &\leftarrow \nu + \triangle \nu \\
\triangle \nu &= -\eta \frac{\partial E_k}{\partial \nu}
\end{aligned} \tag{105}
$$

参数更新 (BP) 反向计算梯度

$$
\begin{aligned}
&\frac{\partial E_k}{\partial \widehat{y}_j^k} = \widehat{y}_j^k - y_j^k \\
&g_j = \frac{\partial E_k}{\partial \alpha_j} = (\widehat{y}_j^k - y_j^k) f'(\alpha_j - \theta_j) \\
&\triangle w_{hj} = -\eta g_j b_h \\
&\triangle \theta_j = \eta g_j \\
&\frac{\partial E_k}{\partial b_h} = \sum_{j=1}^{l} \frac{\partial E_k}{\partial \alpha_j} \frac{\partial \alpha_j}{\partial b_h} = \sum_{j=1}^{l} g_j w_{hj} \\
&e_h = \frac{\partial E_k}{\partial \beta_h} = (\sum_{j=1}^{l} g_j w_{hj}) f'(\beta_h - \gamma_h) \\
&\triangle v_{ih} = -\eta e_h x_i^k \\
&\triangle \gamma_h = \eta e_h
\end{aligned}
\tag{106}
$$

频繁更新参数, 不同样本更新可能不一致, 随机梯度下降 (SGD)

累计 BP

$$
\triangle \nu = -\eta \frac{\partial E}{\partial \nu}, E = \sum_{k=1}^{m} E_k \tag{107}
$$

标准梯度下降, 参数更新不频繁, 累计误差可能下降缓慢. 实际应用中使用小批量 (small batches of) 数据

动量

$$
\begin{aligned}
&\nu \leftarrow \nu + \triangle \nu^{(t)} \\
&\triangle \nu^{(t)} = -\eta \frac{\partial E_k}{\partial \nu} + \alpha \triangle \nu^{(t-1)}
\end{aligned}
\tag{108}
$$

使梯度下降更平滑

局部最小值 解决方法: 多次不同初始化; 模拟退火; 遗传算法

防止过拟合 使用验证集 (早停)

正则化 $E = \frac{1}{m}\sum_{k=1}^{m} E_k + \omega \sum_i \nu_i^2$

Dropout 和 DropConnect

8 统计学习理论

最小描述长度原理 MDL Kolmogorov 复杂度: 一个对象的 Kolmogorov 复杂度是输出为该对象的计算机程序的最短长度

例:“abababababababababababababababab”可用 print('ab'*16),Kolmogorov 复杂度在 Python 中不超过 14

随机性: 一个字符串是随机的，当且仅当每个产生该字符串的计算机程序至少与字符串本身一样长

最小描述长度原理 (MDL): 统计学习任务是找到数据的最短描述

对于一个数据集 D 和假设空间 H, 可表示为:

$$h^* = \arg\min_{h\in H} L(h) + L(D|h) \tag{109}$$

例: 二元分类数据集 $\{(x^{(i)}, y^{(i)})|_{i=1}^{n}\}$, 多个模型 $h_1, ..., h_m$, 有如下几种编码方式:(1) 编码每个 $x^{(i)}$;(2) 编码一些 h_j;(3) 计算 $\widehat{y}^{(k)} = sign(h_j(x^{(i)}))$, 编码集合 $\{i|y^{(i)} \neq \widehat{y}^{(k)}\}$. 最小编码长度的模型最优.

若所有 h_j 和 i 均用固定长度编码,MDL 相当于最小化 0-1loss 的经验风险

若 h_j 定长编码,i 变长编码,MDL 相当于给每组数据加权

若 h_j 变长编码,i 定长编码,MDL 相当于对参数有喜好

MDL 无需概率论的解释，更加灵活