1) 多组不同参数值初始化多个神经网络. 按标准方法训练后. 取其中误差最小的网作为最终参数

2) 模拟退火. 每一步以一定概率接受比当前更差的结果.

3) 随机梯度下降

4. 其他常见神经网络

1) RBF (Radial Basis Function) 径向基函数

$$\varphi(x)=\sum_{i=1}^{p} w_i \rho(x, c_i) \qquad \rho(x, c_i)=e^{-\beta_i \|x-c_i\|^2}$$

确定神经元中心 c_i. 再利用BP算法来确定 w_i, β_i

2) ART (adaptive resonance theory) 自适应谐振理论) 竞争型学习

winner-take-all

3) SOM (self-organizing map) 自组织映射: 每个输出层神经元会计算该样本与自身携带的权向量之间的距离. 距离最近的神经元成为竞争获胜者 best matching unit 最佳匹配单元

§6 支持向量机

1. 感知器模型

$$y=\operatorname{sgn}(\theta \cdot x)=\begin{cases} 1, & \theta x>0 \\ -1, & \theta x<0 \end{cases}$$

$y\theta x>0$ 正确分类

$y\theta x<0$ 错误分类

Loss function: 期望使所有错分样本到超平面之距离之和最小

$$L=\sum_{i=1}^{k} \frac{-y^i \theta \cdot x^i}{\|\theta\|_2} \rightarrow L=-\sum_{i=1}^{k} y^i \theta x^i \qquad \frac{\partial L(\theta)}{\partial \theta}=-\sum_{i=1}^{k} y^i x^i$$

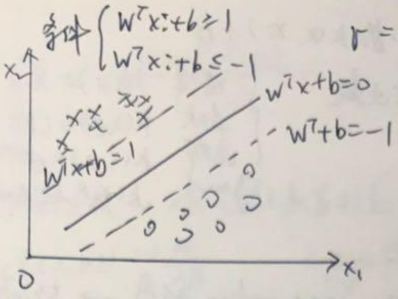
< 线性可分 SVM: 让离超平面较近的点尽可能远离这个点 超平面 (separating hyperplane)
非线性可分 SVM

支持向量 (support vector) 离分割超平面最近的那些点

间隔 (margin): 支持向量表现点到超平面的距离

$W^T x + b = 0$ W—方向 b—位移

$r = \frac{|W^T x + b|}{||W||}$



$r = \frac{2}{||W||}$ (margin)

最大化间隔

$\max_{w,b} \frac{2}{||W||}$

⇕ 对偶化

$\min_{w,b} \frac{1}{2}||W||^2$

$y_i(W^T x_i + b) \geq 1.$

2. 对偶化

特性: 对偶问题的对偶是原问题; 无论原函数是否凸, 对偶问题皆为凸优化;
—— 可给出原始问题下界; 当满足一定条件, 原始问题与对偶问题
是完全等价的.

拉格朗日乘子法. $L(w, b, a) = \frac{1}{2}||W||^2 + \sum_{i=1}^{m} \alpha_i (1 - y_i(W^T x_i + b))$

对w. b偏导为0. $w = \sum_{i=1}^{m} \alpha_i y_i x_i$ $0 = \sum_{i=1}^{m} \alpha_i y_i$

$\Rightarrow \max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i^T x_j$

$f(x) = W^T x + b = \sum_{i=1}^{m} \alpha_i y_i x_i^T x + b$

KKT: $\begin{cases} \alpha_i \geq 0 \\ y_i f(x_i) - 1 \geq 0 \\ \alpha_i (y_i f(x_i) - 1) = 0 \end{cases}$

优化two变量: SMO (Sequential Minimal Opt) 算法
→ 选违背KKT条件程度最大的变量. 使目标函数增长最快. 使所选两变量所对应样本间隔最大

~~3. 核函数 软间隔~~

3. 核函数. 非线性可分SVM 将原始二维空间映射到高维空间

$f(x) = W^T \phi(x + b)$ $\min_{w,b} \frac{1}{2}||W||^2$

$\Rightarrow \max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \phi(x_i)^T \phi(x_j)$

$k(x_i, x_j) = (\phi(x_i), \phi(x_j)) = \phi(x_i)^T \phi(x_j)$

不必计算高维空间中的内积

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j k(x_i, x_j)$$

$$\Rightarrow f(x) = \sum_{i=1}^{m} \alpha_i y_i k(x, x_i) + b$$

核函数只定义正定核函数

- 线性 $k(x,z) = x \cdot z$
- 多项 $k(x,z) = (1 + x \cdot z + r)^d$
- 高斯 $k(x,z) = e^{-r\|x-z\|^2}$
- sigmoid $k(x,z) = \tanh(1 + x \cdot z + r)$

4. 软间隔与正则化

SVM 线性可分效果好. 但对 outlier 敏感 → 软间隔. 松弛因子 ξ_i

ξ_i 越大 → 允许样本偏离

目标函数: 变为 $\frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{m} \xi_i$

KKT 条件 $L(w, b, \xi, \beta, \mu) = \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{m} \xi_i + \sum_{i=1}^{m} \beta_i [1 - \xi_i - y_i(w^T x_i + b)] - \sum_{i=1}^{m} \mu_i \xi_i$

$\min_{w,b,\xi} \max_{\beta,\mu} L(w, b, \xi, \beta, \mu)$ $\Longleftrightarrow$ 对偶优化 $\max_{\beta,\mu} \min_{w,b,\xi} L(w, b, \xi, \beta, \mu)$

求 w, b, ξ 偏导 $\Rightarrow$ $\ell(\beta) = \sum_{i=1}^{m} \beta_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \beta_i \beta_j y_i y_j x_i^T x_j$

更新参数 w^*, b^* $f(x) = \text{sign}(w^* x + b^*)$

此处引到 LR. (LR 损失函数为 sigmoid)

加入松弛因子后. 模型更 robust. 泛化能力强.

Sk-learn 中 SVC (非线性). LinearSVC (线性). OneClassSVM (区分异常)

支持向量也可作回归. SVR 不常用

有约束优化问题: 无约束条件: 梯度下降法 牛顿法 坐标轴下降法 ..

等式约束: 拉格朗日乘子法

不等式约束: KKT