监督学习

1.基本概念

训练误差：训练集的度量误差。
测试误差/泛化误差：对于新输入(样本)的误差期望，一般指测试集的度量误差。
欠拟合：模型不能够在训练集上获得足够低的误差。
过拟合：训练误差和测试误差之间的差异过大，一般是指模型容量大，但数据样本不足。
模型容量：拟合各种函数(输入)的能力。
参数：可以在模型训练中进行更新的参数(如权重矩阵、偏移量)。
超参数：无法通过学习获得，需要自己指定的参数(如学习率、SVM核函数、
正则项L1/L2系数、dropout比率)。
分类问题：标签为类别标签(如2分类、多分类)。
回归问题：标签为连续值。

2.决策树

- C4.5
  样本集合S的熵$H(S)=-\sum_{i=1}^{n}P_ilog_2P_i$
  属性A及被属性A划分成m个的样本集合S_i (0<=i<=m)
  由A划分后剩下的熵的期望$H(S,A)=\sum_{i=0}^{m}\frac{|S_i|}{|S|}H(S_i)$
  A划分的信息增益$gain(S,A)=H(S)-H(S,A)$
  当gain(S,A)的值越大时说明选A的划分时减少了越多的熵，则剩下的熵就越少，划分的纯度也就越高。
  样本由A划分后的分布的熵$splitH(S,A)=-\sum_{i=1}^{m}\frac{|S_i|}{|S|}log_2\frac{|S_i|}{|S|}$
  A划分的信息增益率$gainratio(S,A)=\frac{gain(S,A)}{splitH(S,A)}$
  gain(S,A)表示A划分后的纯度，而splitH(S,A)表示由A划分成m个样本集合的分布越均匀其值越大，
  使gainratio(S,A)值减少，用于对A的划分出现过拟合或是内容不合理时的补偿。
  因此将信息增益率gainratio(S,A)作为A划分数据的信息度量。
  输入：训练样本S，属性集合attrList。
  输出：决策树
  (1)创建根节点cur=root
  (2)if S都属于同一类C，返回cur为叶子并标记为类C
  (3)if attrList为空 or S中的样本数量小于给定的阈值，返回cur为叶子并标记cur为S中出现最多的类
  (4)for each attrList中的属性i，计算gainratio(S,i)
  (5)计算测试属性testAttr = max(gainratio(S,i))
  (6)if testAttr为连续型，找出testAttr的分割阈值
  (7)for each 每个由testAttr划分cur的叶子节点j
  (8)if 叶子节点j对应的样本子集$|Sj|$小于给定的阈值 or $Sj.attrList$为空，转到(3)
  (9)else 令$attrList=Sj.attrList$，转到(4)
  (10)计算每个节点的分类错误，进行剪枝
- CART
  样本集合S，S中各样本的生成概率P_i的基尼指数$gini(S)=\sum_{i=1}^{n}P_i(1-P_i)=1-\sum_{i=1}^{n}P_i^2$
  $gini(P)$表示不确定性，$gini(P)$越小则确定性越高，分类的纯度就越高。
  CART为2叉树，由属性A划分S时分为2类子集，即属于A的1类S_1和不属于A的1类S_2。
  对于多分类问题，需在多个类别(n)中以组合方式枚举其中2类C_n^2。
  由属性A划分S的基尼指数$gini(S,A)=\frac{|S_1|}{|S|}gini(S_1)+\frac{|S_2|}{|S|}gini(S_2)$
  递归创建2叉树CART，递归终止条件：
  (1)叶子都属于同一类别
  (2)样本个数小于给定的阈值
  (3)基尼指数小于给定的阈值
  (4)分类条件与类别的相关程度x^2，当x^2很小时说明分类条件与类别独立，没有必要再分。
  对于回归问题如标签为连续值，则数据集x由CART划分的m个区域$R_1,R_2,\dots,R_m$
  对应每个区间的预测值$\hat{y_1},\hat{y_2},\dots,\hat{y_m}$，定义映射函数$f(x)=\sum_{i=1}^{m}\hat{y_i}I(x\in R_i)$
  平方差公式$\sum_{x_i\in R_i}(y_i-f(x_i))^2$对应于分类问题的基尼指数。
  设属性j的取值tj将输入空间划分为2个区域$R_1(j,tj)=\{x|x^j\le tj\}$， $R_2(j,tj)=\{x|x^j>tj\}$
  设R_1和R_2区间预测的平均值为$\bar{y_1}=\frac{f(x\in R_1)}{|R_1|}$和$\bar{y_2}=\frac{f(x\in R_2)}{|R_2|}$
  目标函数$min(\sum_{x_i\in R_1(j,tj)}(y_i-\bar{y_1})^2+\sum_{x_i\in R_2(j,tj)}(y_i-\bar{y_2})^2)$
  (1)for each 每个属性j
  (2)for each j的每个取值tj
  (3)计算目标函数$min(\sum_{x_i\in R_1(j,tj)}(y_i-\bar{y_1})^2+\sum_{x_i\in R_2(j,tj)}(y_i-\bar{y_2})^2)$
  (4)递归计算R_1和R_2，递归终止条件(每个样本属于1个节点 or 样本数目小于总数的5%)
  (5)计算子节点划分的m个区域$R_1,R_2,\dots,R_m$的映射函数$f(x)=\sum_{i=1}^{m}\hat{y_i}I(x\in R_i)$
- 总结
  优点：
  时间复杂度O(树的深度)
  划分过程中可去除与结果关系不大的属性(可用作对数据的预处理)
  缺点：
  本质为贪心算法(只能得到近似解)
  只能线性分割数据
- 集成决策树
  (1)随机森林：
  m棵树每棵树以有限的样本或是属性(子集)作为输入，但要保证所有的树能包括所有的样本和属性。
  最后以投票的方式或取平均值作为分类结果。
  (2)Boosting：
  m棵树以循环形式处理数据，每次迭代将上次迭代时分类错误的节点加大其权重，再进行处理。
  (3)GBDT：
  m棵树以循环形式处理数据，每次迭代将上次迭代的结果与目标的差距作为新的目标进行残差学习。
  (4)XGBoost：
  m棵树以循环形式处理数据，每次迭代将上次迭代的结果与目标的差距作为新的目标进行残差学习。
  加入L_1或L_2正则项防止过拟合训练数据，以及对目标函数进行1阶、2阶求导来加快收敛。
  设第t回需要进行残差学习的目标为$\sum_{i=1}^{n}f_t(x_i)$，则第t回的预测值为$\sum_{i=1}^{n}\hat{y}^t=\sum_{i=1}^{n}\sum_{j=1}^{t}fj(x_i)$
  满足递推式$\sum_{i=1}^{n}\hat{y}^t=\sum_{i=1}^{n}\hat{y}^{t-1}+\sum_{i=1}^{n}f_t(x_i)$
  则第t回的目标函数$obj^t=\sum_{i=1}^{n}l(y_i,\hat{y}^{t-1}+f_t(x_i))+\Omega(f_t(x_i))$
  将obj^t看作$\hat{y}^{t-1}$的函数进行Tarlor展开，令$g_i=\partial^{\hat{y}^{t-1}}l(y_i,\hat{y}^{t-1})$，$h_i=\partial^2_{\hat{y}^{t-1}}l(y_i,\hat{y}^{t-1})$

$obj^t \approx \sum_{i=1}^{n} g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2 + \Omega(f_t(x_i))$
设当前迭代的叶子数量T，叶子的权重向量w，将输入x映射为其所在叶子的编号q(x)，
L1和L2正则系数为γ和λ
则$f_t(x) = w_{q(x)}$，$\Omega(\sum_{i=1}^{n} f_t(x_i)) = \gamma T + \frac{1}{2}\lambda \sum_{i=1}^{T} w_i^2$
$obj^t \approx \sum_{i=1}^{n} g_i w_{q(x_i)} + \frac{1}{2} h_i w_{q(x_i)}^2 + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$
令$I_j = \{i | q(x_i) = j\}$，$G_j = \sum_{i \in I_j} g_i$，$H_j = \sum_{i \in I_j} h_i$
$obj^t \approx \sum_{j=1}^{T} G_j w_j + \frac{1}{2}(H_j + \lambda) w_j^2 + \gamma T$
则$\min_{w_j} obj^t = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T$
最佳叶子权重$w_j^* = arg \min_{w_j} obj^t = -\frac{G_j}{H_j + \lambda}$

更新树(2叉树)的结构：对已有的叶子进行分割时，为使分割后的obj减少，
定义增益量gain为不分割的obj减去分割后的obj：
$gain = \frac{1}{2}\left(\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right) - \gamma$
这样gain越大则分割后的obj就越小，只需枚举所有叶子的分割点求max(gain)的分割点进行分割。
当gain小于给定的阈值时停止分割。

3.SVM

- 原理：
  为使问题简单一点，假设用一条直线可以将平面上的正负样本点分离开来，如果这样的直线存在2条或者多条，我们希望找出其中最佳的1条直线，把这条直线称为最佳分隔线。
  如何找到这条最佳分隔线呢？就是要让正负样本点都尽可能远离这条直线，也就是让正负样本点到这条直线的距离都最远。方法是用1对平行线把这条最佳分隔线夹在中间，使正负样本各自被分到这对平行线的两侧，把平行线中离正样本近的称为正边界线，离负样本近的为负边界线。
  我们的目标是要这对平行线的间隔最大，相当于让正样本到正边界线的距离最小，同时负样本到负边界线的距离最小。
  平面上的直线可以用斜率和截距这两个参数的方程式表示。那么模型的训练过程就是在训练样本点中计算出满足上述目标的斜率和截距参数，从而确定1条最佳分隔线。
  模型的预测是对于任意1个预测样本点，看它到最佳分隔线的距离(也就是最佳分隔线的法线)的方向，是从正样本到分隔线的距离还是负样本到分隔线的距离，从而确定预测样本点是正样本还是负样本。
  如果一条直线不可以将平面上的正负样本点分离开来，可能要用1条曲线进行分离，同样可以用与这条曲线一样的1对平行曲线，与上述同样的方法进行分类，训练得到的将不在是最佳分隔线，而是1条最佳分隔曲线。区别就在于曲线的方程式表达要更多的参数，也就是在训练中要计算更多的参数而已。同理如果是空间上的样本点，训练得到的就是最佳分隔超平面。
- KKT条件
  (1)最小化目标函数min f(x)和等式约束h(x)=0
  f(x)的极小点x^*满足f(x)和h(x)的梯度在同一直线上
  $\nabla_x f(x^*) = u^* \nabla_x h(x^*)$
  令$L(x,u) = f(x) + uh(x)$
  有$\nabla_x L(x,u) = 0$，$\nabla_u L(x,u) = 0$
  (2)最小化目标函数min f(x)和不等式约束$g(x) \le 0$
  i. 如果f(x)的极小点x^*在$g(x) \le 0$的内部
  有$g(x^*) < 0$ 且 $\nabla_x f(x^*) = 0$
  ii. 如果f(x)的极小点x^*不在$g(x) \le 0$的内部，则x^*满足f(x)的负梯度与g(x)的梯度在同一直线上，
  有 $-\nabla_x f(x^*) = \lambda \nabla_x g(x^*)$， 且$\lambda > 0$，$g(x^*) = 0$
  令$L(x,\lambda) = f(x) + \lambda g(x)$， 对于i：令$\lambda = 0$，有 $\begin{cases} \nabla_x L(x^*,\lambda) = & 0 \\ \lambda \ge & 0 \\ \lambda g(x^*) = & 0 \end{cases}$

  对于$\lambda g(x^*) = 0$这里的λ和$g(x^*)$不能同时为0($\because i, ii$)
  综合(1)(2)有$\min_x f(x)$，s.t：$h(x) = 0$，$g(x) \le 0$
  令$L(x,u,\lambda) = f(x) + uh(x) + \lambda g(x)$，有 $\begin{cases} \nabla_x L(x^*,u,\lambda) = & 0 \\ \lambda \ge & 0 \\ \lambda g(x^*) = & 0 \end{cases}$

  令$p(x) = \max_{u,\lambda} L(x,u,\lambda) = \begin{cases} f(x) & constraints \\ +\infty & else \end{cases}$
  则$p^* = \min_x p(x) = \min_x f(x)$
  令$d(x) = \min_x L(x,u,\lambda)$
  则$d^* = \max_{u,\lambda} d(x)$

  弱对偶：$d^* \le p^*$ ($\because$最小里面的最大总是小于等于最大里面的最小)
  强对偶：$d^* = p^*$
- 分类问题
  设最佳分割线的法线单位向量w，原点到最佳分割线的距离b
  则正样本向量x_+需满足$x_+ \cdot w \ge b$， 负样本向量x_-需满足$x_- \cdot w \le b$，
  即$\begin{cases} x_+ \cdot w - b \ge & 0 \\ x_- \cdot w - b \le & 0 \end{cases}$
  假设正负样本与最佳分割线存在单位1的间隔，
  有$\begin{cases} x_+ \cdot w - b \ge & 1 \\ x_- \cdot w - b \le & -1 \end{cases}$
  正负样本都乘上各自的标签，得到约束条件$y(x \cdot w - b) \ge 1$
  设松弛变量$\varepsilon_i \ge 0$，对约束加入松弛变量$y_i(x_i \cdot w - b) \ge 1 - \varepsilon_i$
  设正负样本向量(点)x_+和x_-都在各自的间隔边界上，
  有$\begin{cases} x_+ \cdot w - b = & 1 \\ x_- \cdot w - b = & -1 \end{cases}$
  则最大间隔$max((x_+ - x_-) \cdot \frac{w}{\|w\|}) = max(\frac{2}{\|w\|})$
  相当于求$\min_{w,b} \frac{1}{2}||w||^2$ ($||w||^2$包括法线单位向量w和长度b)

还要限制松弛变量εi，设惩罚系数$C>0$，样本总数n
则$\min\limits_{w,b,\varepsilon_i} \frac{1}{2}||w||^2+C\sum_{i=1}^{n}\varepsilon i$
得到优化问题$\min\limits_{w,b,\varepsilon_i} \frac{1}{2}||w||^2+C\sum_{i=1}^{n}\varepsilon i$

$$s.t\begin{cases}1-\varepsilon_i-y_i(x_i\cdot w-b)\leq & 0\\ -\varepsilon_i\leq & 0\end{cases}$$

其中$C>0$，令满足KKT条件的$\alpha_i,\beta_i\geq 0$，
有$L=\frac{1}{2}w^Tw+C\sum_{i=1}^{n}\varepsilon i+\sum_{i=1}^{n}\alpha_i[1-\varepsilon_i-y_i(x_i\cdot w-b)]-\sum_{i=1}^{n}\beta_i\varepsilon i$
原问题$p^*=\min\limits_{w,b,\varepsilon_i}\max\limits_{\alpha_i,\beta_i} L$，求其对偶问题$d^*=\max\limits_{\alpha_i,\beta_i}\min\limits_{w,b,\varepsilon_i} L$

$$\begin{cases}\frac{\partial L}{\partial w}= & 0\\ \frac{\partial L}{\partial b}= & 0\\ \frac{\partial L}{\partial \varepsilon_i}= & 0\end{cases}$$

得到L只是α_i的函数，
即$d^*=\max\limits_{\alpha_i}{}_n L=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_jy_iy_jx_i^Tx_j$

$$s.t\begin{cases}\sum\limits_{i=1}\alpha_iy_i=0\\ 0\leq\alpha_i\leq C\end{cases}$$

满足KKT条件的最优解$\alpha_i^*,\varepsilon_i^*,w^*,b^*$，

$$有\begin{cases}\alpha_i^*[1-\varepsilon_i^*-y_i(x_i\cdot w^*-b^*)]= & 0\\ \beta_i\varepsilon_i^*=(C-\alpha_i^*)\varepsilon_i^*= & 0\end{cases}$$

当$\alpha_i^*=0$，有$\varepsilon_i^*=0$，则$y_i(x_i\cdot w^*-b^*)=1$，
说明x_i不是支持向量
当$0<\alpha_i^*<C$，有$\varepsilon_i^*=0$，则$y_i(x_i\cdot w^*-b^*)=1$，
说明x_i是支持向量(在间隔边界上)
当$\alpha_i^*=C$，有$\varepsilon_i^*=0$，则$1-y_i(x_i\cdot w^*-b^*)=\varepsilon_i^*>0$，
说明$y_i(x_i\cdot w^*-b^*)<1$(x_i在间隔边界内部)
对于不满足KKT条件的样本x_i，进行更新α_i,w,b
当$y_i(x_i\cdot w-b)>1$，有$\alpha_i>0$
当$y_i(x_i\cdot w-b)<1$，有$\alpha_i<C$
当$y_i(x_i\cdot w-b)=1$，有$\alpha_i=0$ 或 $\alpha_i=C$
每次选择2个α_i,α_j更新(保证α_i线性独立)
令迭代中上次预测值$u_i=x_i\cdot w-b$，预测值与真实值之差$E_i=u_i-y_i$
while(存在不满足KKT条件的α_i)
设$\alpha_1^{new},\alpha_2^{new}$为本次迭代需要更新的值
为快速收敛贪心地在满足KKT条件中找$max(|E_i-Ej|)$的αj
令$\alpha_1^{old}=\alpha_i,y_1=y_i,\alpha_2^{old}=\alpha j,y_2=yj$
则满足$\alpha_1^{new}y_1+\alpha_2^{new}y_2=\alpha_1^{old}y_1+\alpha_2^{old}y_2=-\sum_{i=3}^{n}\alpha_iy_i$，两边乘$y_1$得
$\alpha_1^{new}+\alpha_2^{new}y_2y_1=\alpha_1^{old}+\alpha_2^{old}y_2y_1=-y_1\sum_{i=3}^{n}\alpha_iy_i$
令$s=y_1y_2,t=-y_1\sum_{i=3}^{n}\alpha_iy_i$，有
$\alpha_1^{new}+\alpha_2^{new}s=\alpha_1^{old}+\alpha_2^{old}s=t,\quad \alpha_1^{new}=t-\alpha_2^{new}s$
求α_2^{new}的上下界L,H：

$$\begin{cases}L=max(0,\alpha_2^{old}-\alpha_1^{old}),\quad H=min(C,C+\alpha_2^{old}-\alpha_1^{old}) & if(y_1=y_2)\\ L=max(0,\alpha_1^{old}+\alpha_2^{old}-C),\quad H=min(C,\alpha_1^{old}+\alpha_2^{old}) & if(y_1=y_2)\end{cases}$$

$$\begin{aligned}V_i=\sum_{j=3}^{n}y_j\alpha_jx_j^Tx_i&=\sum_{k=1}^{n}y_k\alpha_kx_k^Tx_i-y_1\alpha_1^{old}x_1^Tx_i-y_2\alpha_2^{old}x_2^Tx_i-b+b\\ &=w\cdot x_i-b+b-y_1\alpha_1^{old}x_1^Tx_i-y_2\alpha_2^{old}x_2^Tx_i\\ &=u_i+b-y_1\alpha_1^{old}x_1^Tx_i-y_2\alpha_2^{old}x_2^Tx_i\end{aligned}$$

$$\begin{aligned}d^*=\max_{\alpha_i}L=\min_{\alpha_i}-L&=\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_jy_iy_jx_i^Tx_j-\sum_{i=1}^{n}\alpha_i\\ &=\frac{1}{2}x_1^Tx_1\alpha_1^{new^2}+\frac{1}{2}x_2^Tx_2\alpha_2^{new^2}+y_1y_2\alpha_1^{new}\alpha_2^{new}x_1^Tx_2+y_1\alpha_1^{new}V_1+y_2\alpha_2^{new}V_2\\ &\quad-\alpha_1^{new}-\alpha_2^{new}+const(\alpha_3\ldots\alpha_n)\end{aligned}$$

将d^*表示为只含α_2^{new}的函数，并对其求导：

$$\begin{aligned}d^*&=\frac{1}{2}x_1^Tx_1(t-s\alpha_2^{new})^2+\frac{1}{2}x_2^Tx_2\alpha_2^{new^2}+s(t-s\alpha_2^{new})\alpha_2^{new}x_1^Tx_2+y_1(t-s\alpha_2^{new})V_1\\ &\quad+y_2\alpha_2^{new}V_2-t+s\alpha_2^{new}-\alpha_2^{new}+const(\alpha_3\ldots\alpha_n)\end{aligned}$$

$$\begin{aligned}\frac{\partial d^*}{\partial\alpha_2^{new}}&=-sx_1^Tx_1(t-s\alpha_2^{new})+x_2^Tx_2\alpha_2^{new}-x_1^Tx_2\alpha_2^{new}+sx_1^Tx_2(t-s\alpha_2^{new})-y_2V_1\\ &\quad+s+y_2V_2-1=0\end{aligned}$$

$$\alpha_2^{new}=\alpha_2^{old}+\frac{y_2(E_1-E_2)}{x_1^Tx_1+x_2^Tx_2-2x_1^Tx_2}$$

考虑α_2^{new}的范围有

^{new}

$$\alpha_2^{new}=\begin{cases}H & if(\alpha_2^{new}>H)\\ \alpha_2^{new} & if(L\le\alpha_2^{new}\le H)\\ L & if(\alpha_2^{new}<L)\end{cases}$$

$\alpha_1^{new}=\alpha_1^{old}+y_1y_2(\alpha_2^{old}-\alpha_2^{new})$
更新参数$w^{new}=\sum_{i=1}^{n}\alpha_iy_ix_i$
令$y_1\approx w^{new}\cdot x_1-b_1^{new}$有$E_1=u_1-y_1\approx w\cdot x_1-b^{old}-(w^{new}\cdot x_1-b_1^{new})$
$b_1^{new}\approx b^{old}+E_1+x_1[\sum_{i=3}^{n}x_iy_i\alpha_i+x_1y_1\alpha_1^{new}+x_2y_2\alpha_2^{new}$
$-(\sum_{i=3}^{n}x_iy_i\alpha_i+x_1y_1\alpha_1^{old}+x_2y_2\alpha_2^{old})]$
$\approx b^{old}+E_1+y_1(\alpha_1^{new}-\alpha_1^{old})x_1^Tx_1+y_2(\alpha_2^{new}-\alpha_2^{old})x_1^Tx_2$

同样：$b_2^{new}\approx b^{old}+E_2+y_1(\alpha_1^{new}-\alpha_1^{old})x_2^Tx_1+y_2(\alpha_2^{new}-\alpha_2^{old})x_2^Tx_2$
更新参数
$$b^{new}=\begin{cases}b_1^{new} & if(0<\alpha_1^{new}<C)\\ b_2^{new} & if(0<\alpha_2^{new}<C)\\ \dfrac{b_1^{new}+b_2^{new}}{2} & else\end{cases}$$

最后由求得的参数w^*,b^*确定最佳分界线的方程，用来预测新样本$f(x)=sign(w^*\cdot x-b^*)$

- 非线性
核函数例
$$K(x_1,x_2)=\begin{cases}(x_1\cdot x_2+R)^d, & i.e:(x_1\cdot x_2+1)^2\\ exp(-\dfrac{|x_1-x_2|^2}{2\sigma}), & \sigma>0\\ tanh(a(x_1\cdot x_2)+b), & a,b=const\end{cases}$$

优化问题改为$d^*=\max_{\alpha_i}L=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_jy_iy_jK(x_i,x_j)$
$$s.t\begin{cases}\sum_{i=1}\alpha_iy_i=0\\ 0\le\alpha_i\le C\end{cases}$$

预测新样本$f(x)=sign(\sum_{i=1}^{n}\alpha_i^*y_iK(x,x_i)-b^*)$

- 回归问题
数据集$\{x_i,t_i\}$，t_i为连续的值，设$f(x)=w\cdot x-b$，误差$\epsilon>0$的函数
$$E_i(f(x_i)-t_i)=\begin{cases}0 & if(|f(x_i)-t_i|<\epsilon)\\ |f(x_i)-t_i|-\epsilon & else\end{cases}$$
加入松弛变量约束$\sigma_i,\hat{\sigma}_i$,
$$有\begin{cases}t_i-f(x_i)-\sigma_i\le\epsilon & if(\sigma_i>0\quad and\quad\hat{\sigma}i=0\quad and\quad t_i>f(x_i)+\epsilon)\\ f(x_i)-t_i-\hat{\sigma}_i\le\epsilon & if(\hat{\sigma}_i>0\quad and\quad\sigma_i=0\quad and\quad t_i<f(x_i)-\epsilon)\end{cases}$$
则优化问题
$\min_{w,b,\sigma_i,\hat{\sigma}_i}\frac{1}{2}||w||^2+C\sum_{i=1}^{n}(\sigma_i+\hat{\sigma}_i)$
$$s.t\begin{cases}t_i-f(x_i)-\sigma_i-\epsilon\le & 0\\ f(x_i)-t_i-\hat{\sigma}_i-\epsilon\le & 0\\ -\sigma_i\le & 0\\ -\hat{\sigma}_i\le & 0\end{cases}$$

其中$C>0$，令满足KKT条件的$\alpha_i,\hat{\alpha}_i,\beta_i,\hat{\beta}_i\ge0$，有
$L=\frac{1}{2}w^Tw+C\sum_{i=1}^{n}(\sigma_i+\hat{\sigma}_i)+\sum_{i=1}^{n}\alpha_i[t_i-(w\cdot x_i-b)-\sigma_i-\epsilon]$
$+\sum_{i=1}^{n}\hat{\alpha}_i[(w\cdot x_i-b)-t_i-\hat{\sigma}_i-\epsilon]-\sum_{i=1}^{n}\beta_i\sigma_i-\sum_{i=1}^{n}\hat{\beta}_i\hat{\sigma}_i$

原问题$p^*=\min_{w,b,\sigma_i,\hat{\sigma}_i}\max_{\alpha_i,\hat{\alpha}_i,\beta_i,\hat{\beta}_i}L$，求其对偶问题$d^*=\max_{\alpha_i,\hat{\alpha}_i,\beta_i,\hat{\beta}_i}\min_{w,b,\sigma_i,\hat{\sigma}_i}L$
$$\begin{cases}\dfrac{\partial L}{\partial w}= & 0\\ \dfrac{\partial L}{\partial b}= & 0\\ \dfrac{\partial L}{\partial\sigma_i}= & 0\\ \dfrac{\partial L}{\partial\hat{\sigma}_i}= & 0\end{cases}$$

得到L只是$\alpha_i,\hat{\alpha}_i$的函数，即
$d^*=\max_{\alpha_i,\hat{\alpha}_i}L=\sum_{i=1}^{n}(\alpha_i-\hat{\alpha}_i)t_i-\epsilon\sum_{i=1}^{n}(\alpha_i+\hat{\alpha}_i)-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i-\hat{\alpha}_i)(\alpha_j-\hat{\alpha}^j)x_i^Tx_j$
$$s.t\begin{cases}0\le\alpha_i\le C\\ 0\le\hat{\alpha}^j\le C\end{cases}$$

满足KKT条件的最优解$\alpha_i^*, \hat{\alpha i}^*, \sigma_i^*, \hat{\sigma i}^*, w^*, b^*$，

有$\begin{cases} \alpha_i^*[t_i - (w^* \cdot x_i - b^*) - \sigma_i^* - \epsilon] = 0 \\ \hat{\alpha i}^*[(w^* \cdot x_i - b^*) - t_i - \hat{\sigma i}^* - \epsilon] = 0 \\ (C - \alpha_i^*)\sigma_i^* = 0 \\ (C - \hat{\alpha i}^*)\hat{\sigma i}^* = 0 \end{cases}$

当$\alpha_i^* = C, \sigma_i^* > 0$，有$t_i > (w^* \cdot x_i - b^*) + \epsilon$， 说明$t_i$在$\epsilon$定义的管道上边界的上方。
当$\hat{\alpha i}^* = C, \hat{\sigma i}^* > 0$，有$t_i < (w^* \cdot x_i - b^*) - \epsilon$， 说明$t_i$在$\epsilon$定义的管道下边界的下方。
当$\alpha_i^* = \hat{\alpha i}^* = 0$，有$(w^* \cdot x_i - b^*) - \epsilon \le t_i \le (w^* \cdot x_i - b^*) + \epsilon$， $\sigma_i^* = \hat{\sigma i}^* = 0$，
说明t_i在ϵ定义的管道内部。
当$0 < \alpha_i^*, \hat{\alpha i}^* < C$，有$\sigma_i^* = \hat{\sigma i}^* = 0$，$t_i = (w^* \cdot x_i - b^*) + \epsilon \quad or \quad t_i = (w^* \cdot x_i - b^*) - \epsilon$，
说明t_i在ϵ定义的管道上边界线或下边界线上。
对于不满足KKT条件的样本x_i，进行更新$\alpha_i, \hat{\alpha i}, w, b$
当$t_i > (w \cdot x_i - b) + \epsilon$，有$0 \le \alpha_i < C$
当$t_i < (w^* \cdot x_i - b^*) - \epsilon$，有$0 \le \hat{\alpha i} < C$
当$(w \cdot x_i - b) - \epsilon \le t_i \le (w \cdot x_i - b) + \epsilon$， 有$0 < \alpha_i, \hat{\alpha i} \le C$
当$t_i = (w \cdot x_i - b) + \epsilon \quad or \quad t_i = (w \cdot x_i - b) - \epsilon$， 有$\alpha_i, \hat{\alpha i} = 0 \quad or \quad \alpha_i, \hat{\alpha i} = C$
更新$\alpha_i, \hat{\alpha i}, w, b$时与分类问题同样每次选1对$(\alpha_i, \hat{\alpha i})$进行更新。
最后由求得的参数$\alpha_i^*, \hat{\alpha i}^*, w^*, b^*$确定最佳分界线的方程，用来预测新样本
$f(x) = sign(w^* \cdot x - b^*) = sign(\sum_{i=1}^{n}(\alpha_i^* - \hat{\alpha i}^*)x_i^T x - b^*)$
如采用非线性则$f(x) = sign(\sum_{i=1}^{n}(\alpha_i^* - \hat{\alpha i}^*)K(x, x_i) - b^*)$

- 总结
  优点：
  对非线性的分类回归问题及高维度的数据有效。
  由$\alpha_i = 0$作为支持向量满足稀疏性，无需依赖全部数据作决策。
  缺点：
  计算量大，对缺失数据敏感。

4.神经网络

- 原理
  多项式可以用来拟合任意连续函数，例如：
  2项式定理：$(x+y)^n = \sum_{k=0}^{n}\binom{n}{k}x^k y^{n-k}$，两边$x$微分后乘$\frac{x}{n}$有
  $x(x+y)^{n-1} = \sum_{k=0}^{n}\binom{n}{k}\frac{k}{n}x^k y^{n-k}$， 令$y = 1-x$， 有 $x = \sum_{k=0}^{n}\binom{n}{k}\frac{k}{n}x^k(1-x)^{n-k}$
  再次两边x微分后乘$\frac{x}{n}$，且令$y = 1-x$， 有
  $x^2 + \frac{x(1-x)}{n} = \sum_{k=0}^{n}\binom{n}{k}(\frac{k}{n})^2 x^k(1-x)^{n-k}$， 当$n = +\infty$时有$x^2 = \sum_{k=0}^{n}\binom{n}{k}(\frac{k}{n})^2 x^k(1-x)^{n-k}$
  可以看出如果将x作为参数，等式右边含$\frac{x}{n}$的项看作函数形式的话有
  $B_n(x, f) = \sum_{k=0}^{n}\binom{n}{k}f(\frac{k}{n})x^k(1-x)^{n-k}$， 当$n = +\infty$时有$\lim_{n\to\infty} B_n(x, f) = f(x)$
  $B_n(x, f)$可解释为在区间$[0,1]$中任取n个数，令落在区间$[0, x]$的个数k，落在区间$(x, 1]$的个数$n-k$
  则落在区间$[0, x]$的个数k的概率为$P_k = \binom{n}{k}x^k(1-x)^{n-k}$，$B_n(x, f)$就是$f(\frac{k}{n})$的期望
  $B_n(x, f) = E^{P_k}[f(\frac{k}{n})] = \sum_{k=0}^{n}\binom{n}{k}f(\frac{k}{n})x^k(1-x)^{n-k}$
  定理：$f(x)$为m维空间$C(I_m), I_m = [0,1]^m$上的连续函数，则$\lim_{n\to\infty} B_n(x, f) = f(x)$
  证明：令$|f(x) - B_n(x, f)| = \sum_{k=0}^{n}\binom{n}{k}|f(x) - f(\frac{k}{n})|x^k(1-x)^{n-k}$
  因为$f(x)$在区间$[0,1]$上有界，则存在M使$|f(x) < M|$， 则
  $|f(x) - f(\frac{k}{n})| < 2M, \quad x, \frac{k}{n} \in [0,1]$
  因为$f(x)$在区间$[0,1]$上连续，则对于任意$\epsilon > 0$， 都存在$\delta(\epsilon)$使
  $|f(x) - f(\frac{k}{n})| < \epsilon, \quad if(|x - \frac{k}{n}| < \delta(\epsilon))$
  考虑满足$|x - \frac{k}{n}| < \delta(\epsilon)$的上界$\epsilon$， 令其可能出现的总数$S_1$有
  $S_1 = \epsilon\sum^{|x-\frac{k}{n}|<\delta(\epsilon)}\binom{n}{k}x^k(1-x)^{n-k} \le \epsilon\sum_{k=0}^{n}\binom{n}{k}x^k(1-x)^{n-k} = \epsilon$
  考虑$|x - \frac{k}{n}| \ge \delta(\epsilon)$的上界$2M$， 令其可能出现的总数$S_2$有
  $S_2 = 2M\sum^{|x-\frac{k}{n}|\ge\delta(\epsilon)}\binom{n}{k}x^k(1-x)^{n-k}$， 因为$\frac{(x-\frac{k}{n})^2}{\delta(\epsilon)^2} \ge 1$， 有
  $S_2 \le \frac{2M}{\delta(\epsilon)^2}\sum_{k=0}^{n}\binom{n}{k}(x-\frac{k}{n})^2 x^k(1-x)^{n-k} = \frac{2M}{\delta(\epsilon)^2}[x^2 B_n(x,1) - 2xB_n(x,x) + B_n(x,x^2)]$
  $= \frac{2M}{\delta(\epsilon)^2}[x^2 - 2x^2 + x^2 + \frac{x(1-x)}{n}] \le \frac{2M}{4\delta(\epsilon)^2 n}$

  $\therefore |f(x) - B_n(x, f)| \le S_1 + S_2 \le \epsilon + \frac{2M}{4\delta(\epsilon)^2 n}$， 当$n \to \infty$时有$|f(x) - B_n(x, f)| \le \epsilon$
  多层前馈神经网络
  $\begin{cases} h_1 = \quad \sigma(w_1 \cdot x + b_1) \\ h_2 = \quad \sigma(w_2 \cdot h_1 + b_2) \\ \dots \\ \hat{y} = \quad \sigma(w_k \cdot h^{k-1} + bk) \end{cases}$

  可以看出每层都是通过前1层的结果h与当前层的权重w进行线性组合得出结果，
  最终的结果$\hat{y}$可以看作与所有层的权重w线性组合的结果，即可以表示为
  $\hat{y} = \sum_{i=1}^{k}\alpha i\sigma(\sum_{j=1}^{n}wij + bi)$， 令$\alpha k \to f(\frac{k}{n})$， $\sigma(\sum_{j=1}^{n}wkj + bk) \to Pk$有
  $\hat{y} = Bn(x, f) = f(x), \quad if(n \to \infty)$
- 误差反向传播
  设前馈神经网络(向量成分表示)输入层56维$x_{\cdot\times 56}$，隐含层2维$h_{\cdot\times 2}$，输出层1维$\hat{y_{\cdot\times 1}}$，
  则输入层与隐含层的权重$w^1_{56\times 2}$，隐含层与输出层的权重$w^2_{2\times 1}$，
  有$\begin{cases} h_{\cdot\times 2} = \quad \sigma(x_{\cdot\times 56} \cdot w^1_{56\times 2}) \\ \hat{y_{\cdot\times 1}} = \quad \sigma(h_{\cdot\times 2} \cdot w^2_{2\times 1}) \end{cases}$
  令目标函数$E_{\cdot\times 1} = \frac{1}{2}(y_{\cdot\times 1} - \hat{y_{\cdot\times 1}})^2$， 学习率$\eta$
  输入层与隐含层$w^1_{56\times 2}$的调整

$\frac{\partial E_{\cdot\times 1}}{}$

$$
\begin{aligned}
\overline{\partial w^{1}_{56\times 2}} &= -(y_{\cdot\times 1}-\hat{y^{\wedge}}_{\cdot\times 1})\hat{y^{\wedge}}'_{\cdot\times 1} \\
&= -(y_{\cdot\times 1}-\hat{y^{\wedge}}_{\cdot\times 1})\frac{\partial \hat{y^{\wedge}}_{\cdot\times 1}}{\partial h_{\cdot\times 2}}\frac{\partial h_{\cdot\times 2}}{\partial x_{\cdot\times 56}\cdot w^{1}_{56\times 2}}\frac{\partial x_{\cdot\times 56}\cdot w^{1}_{56\times 2}}{w^{1}_{56\times 2}} \\
&= -(y_{\cdot\times 1}-\hat{y^{\wedge}}_{\cdot\times 1})\begin{bmatrix} w^{2}_{2\times 1}[0] \\ w^{2}_{2\times 1}[1] \end{bmatrix}\odot\sigma(x_{\cdot\times 56}\cdot w^{1}_{56\times 2})'\odot x_{\cdot\times 56} \\
&= -(y_{\cdot\times 1}-\hat{y^{\wedge}}_{\cdot\times 1})\cdot w^{2T}_{1\times 2}\odot\sigma(x_{\cdot\times 56}\cdot w^{1}_{56\times 2})'\odot x_{\cdot\times 56}
\end{aligned}
$$

可以看出在误差反向传播的过程中也是在进行权重的线性组合，则调整权重为

$\triangle w^{1}_{56\times 2}=-\eta^{\frac{\partial E_{\cdot\times 1}}{\partial w^{1}_{56\times 2}}}=\eta(y_{\cdot\times 1}-\hat{y^{\wedge}}_{\cdot\times 1})\cdot w^{2T}_{1\times 2}\odot\sigma(x_{\cdot\times 56}\cdot w^{1}_{56\times 2})'\odot x_{\cdot\times 56}$

隐含层与输出层$w^{2}_{2\times 1}$的调整

$\frac{\partial E_{\cdot\times 1}}{\partial w^{2}_{2\times 1}}=-(y_{\cdot\times 1}-\hat{y^{\wedge}}_{\cdot\times 1})\hat{y^{\wedge}}'_{\cdot\times 1}=-(y_{\cdot\times 1}-\hat{y^{\wedge}}_{\cdot\times 1})\frac{\partial \hat{y^{\wedge}}_{\cdot\times 1}}{\partial w^{2}_{2\times 1}}=-(y_{\cdot\times 1}-\hat{y^{\wedge}}_{\cdot\times 1})\odot h_{\cdot\times 2}$

则调整权重为$\triangle w^{2}_{2\times 1}=-\eta^{\frac{\partial E_{\cdot\times 1}}{\partial w^{2}_{2\times 1}}}=\eta(y_{\cdot\times 1}-\hat{y^{\wedge}}_{\cdot\times 1})\odot h_{\cdot\times 2}$

- 批量归一化

  例如σ激活函数远离原点的两侧存在平坦区域，如果在误差反向传播过程中某隐含层的梯度来自于平坦区域，则此梯度几乎得不到更新，而且根据求导的链式法则，每层的梯度都来自于前1层的输入，那么在此层之前各层的梯度也都得不到更新，这就是梯度消失的问题。

  批量归一化就是要让每层的梯度都来自于原点附近的非平坦区域，从而防止梯度消失的问题。

  令在每层之间加入1个归一化层$\hat{h_i}$，设h_i表示$\hat{h_i}$的后1层，$\bar{h_i}$表示h_i的均值，

  $var(h_i)$表示h_i的方差，$\epsilon>0$表示防止$var(h_i)=0$的很小的正数，有

  $\hat{h_i}=\frac{h_i-\bar{h_i}}{\sqrt{var(h_i)+\epsilon}}$，为使其移动至原点附近，加入参数$\beta,\gamma$

  使$\hat{h_i}=\beta\hat{h_i}+\gamma$，$\beta,\gamma$可通过训练自适应地得到。

- 自归一化

  设H为某隐含层，其输入x，输出y，则$y=H(x)$，$\frac{\partial y}{\partial x}=\nabla xH$

  梯度消失的问题可理解为x很大的变化导致了y很小的变化，也就是指H在误差反向传播过程中缩小了梯度，使得y的方差小于x的方差。同样如果x很小的变化导致了y很大的变化，也就是H在误差反向传播过程中放大了梯度，使得y的方差大于x的方差，从而可能导致梯度爆炸。

  自归一化就是要让H能够保证其输入x和输出y的方差稳定不变，从而防止梯度消失或爆炸的问题。

  令激活函数

  $$
  selu(x)=\begin{cases}\lambda x & if(x>0,\lambda>1)\\ \lambda\alpha(e^{x}-1) & if(x\le 0,\lambda>1)\end{cases}
  $$

  可以看出原点右侧由于$\lambda>1$，可以扩大方差。原点左侧为衰减函数则减小方差。

  调整λ,α可使整体方差稳定不变。

  对于输入x为正态分布，可求出$\lambda=1.0506,\alpha=1.67326$

  如果输入x不为正态分布，可初始化权重w为均值0，方差1的标准正态分布。

  此时(0,1)为均值方差在相空间中的不动点，从而使整体方差稳定不变。

- 优化器

  令学习率η，位置θ，速度v，则$vt=\theta_{t-1}-\theta_t=\eta\nabla E(\theta_{t-1})$

  梯度下降$\theta_t=\theta_{t-1}-\eta\nabla E(\theta_{t-1})=\theta_{t-1}-v_t$

  局部梯度的反方向不一定是函数整体下降方向，例如隧道型曲面。因此需要加入动量γ用来保持上次速度的比例$v_t=\gamma v_{t-1}+\eta\nabla E(\theta_{t-1})$，从而加速收敛。但这样会积累函数整体下降方向的速度，导致在到达最低点却无法停下来。需要预先判断下1步的位置$\theta_t=\theta_{t-1}-\gamma v_{t-1}$来做出调整。

  所以梯度下降可改成

  $$
  \begin{cases}v_t=\gamma v_{t-1}+\eta\nabla E(\theta_{t-1}-\gamma v_{t-1})\\ \theta_t=\theta_{t-1}-v_t\end{cases}
  $$

  对于给定的学习率η并非都适合所有参数，例如某些参数更新频率小则需要很大的学习率，某些参数更新频率大则需要很小的学习率。因此需要不同的学习率来对应不同的参数且满足所有参数的学习率都要衰减，希望学习率可以通过训练自适应地调整。所以需要记录当前时刻之前更新了多少参数。

  考虑随时刻t的参数$g^{2}_{t,i}=v^{2}_{t,i}$历史平方和的移动平均值$E[g^2]_{t,i}=\gamma E[g^2]_{t-1,i}+(1-\gamma)g^{2}_{t,i}$

  一般取$\gamma=0.9$，则有$\theta_{t,i}=\theta_{t-1,i}-\frac{\eta}{\sqrt{E[g^2]_{t-1,i}+\epsilon}}g_{t-1,i}$

  为使上面的第2项具备θ单位，令$\triangle\theta_{t,i}=\theta_{t,i}-\theta_{t-1,i}$　$E[\triangle\theta^2]_{t,i}=\gamma E[\triangle\theta^2]_{t-1,i}+(1-\gamma)\triangle\theta^{2}_{t,i}$

  此时的学习率η没用了，所以将其替换成 $\theta_{t,i}=\theta_{t-1,i}-\frac{\sqrt{E[\triangle\theta^2]_{t-1,i}}}{\sqrt{E[g^2]_{t-1,i}+\epsilon}}g_{t-1,i}$

  结合动量和自适应学习率得

  $$
  \begin{cases}v_{t,i}=\gamma v_{t-1,i}+\frac{\sqrt{E[\triangle\theta^2]_{t-1,i}}}{\sqrt{E[g^2]_{t-1,i}+\epsilon}}(\nabla E(\theta_{t-1,i}-\gamma v_{t-1,i}))^2\\ \theta_{t,i}=\theta_{t-1,i}-v_{t,i}\end{cases}
  $$

- dropout

  通过在训练实例上随机忽略1些隐含节点，可认为产生不同的模型。如隐含节点个数n，以0.5的概率随机忽略每个隐含节点，等效于从2^n个不同模型中随机采用1个。

  因为权重共享，每个模型都被正则化，同时提高了单个隐含节点的能力，也就提高了整体模型的泛化能力(因为协作方式可能过于复杂，在新的数据集上效果不佳)，从而可防止过拟合训练数据。

5.正则项

L_1正则如$w^{*}_{j}=arg\min_{w_j}\frac{1}{2n}\sum_{i=1}^{n}\sqrt{(y_i-x_i\cdot w_{ij})^2}+\lambda\sum_{i=1}^{n}|w_{ij}+w_{i0}|$

w_{i0}为偏移量，令常数$C_i=-w_{i0}$，上式的最优解w^{*}_{j}满足KKT条件，有$\lambda=0$，$\sum_{i=1}^{n}|w^{*}_{ij}-C_i|=0$

如2维空间中有$w^{*}_{1}+w^{*}_{2}=C$，很大概率使得$w^{*}_{1}=0$或$w^{*}_{2}=0$，从而产生稀疏矩阵。

L1会趋向于产生少量的特征，可用作特征选择。

L_2正则如$w^{*}_{j}=arg\min_{w_j}\frac{1}{2n}\sum_{i=1}^{n}\sqrt{(y_i-x_i\cdot w_{ij})^2}+\lambda\sum_{i=1}^{n}w^{2}_{ij}$

L_2正则限制了展开多项式中的平方项的影响程度，使模型容量减少。

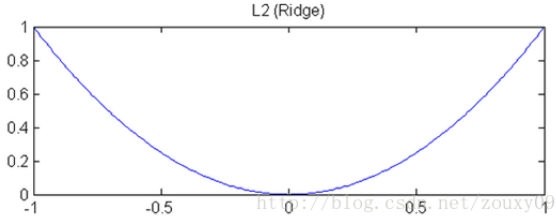
将L_2正则公式简写成$L=E+\frac{\lambda}{2}w^2$，对w求导$\frac{\partial L}{\partial w}=\frac{\partial E}{\partial w}+\lambda w^{*}=0$

得$w^{*}=-\frac{1}{\lambda}\frac{\partial E}{\partial w}$，可以看出如果有很大的权重只可能出现在梯度的反方向，即得到最优解的方向。

因此可防止模型利用权重去拟合错误的样本(如异常值)。

对于大小相同的输入如$1+0=1$，$0.5+0.5=1$，由于正则项$1^2+0^2=1$，$0.5^2+0.5^2=0.5$

可知得到大的值1会被惩罚，所以模型会使权重平摊而减少。
在权重较小时下面的L与w的图像可知输出权重的变化量很小，模型趋近于平稳。



对比L1会趋向于产生少量的特征，L2则会很平均地选择权重较小的特征。
因此在所有特征中只有少数特征起重要作用的情况下，选择L1比较合适，因为它能自动选择特征。
如果所有特征中，大部分特征都能起作用，而且起的作用很平均时选择L2合适。

6.评价指标

对于2分类问题，准确率 := (正判正+负判负) / (正判正+负判负+正判负+负判正)
一般都采用准确率作为模型的评价指标，然而在某些情况下准确率并不都能够反应出模型存在的问题。比如自动驾驶需要去训练识别路标，如果将通行标示识别成停止，那么停下来可能会阻碍交通。但是反过来将停止标示识别成通行，那么通行的话就可能造成交通事故。
这两种情况的后果是天壤之别，对比于把其他的标示识别成停止，更不希望把停止标示识别成其他标示。
换作2分类问题，如果停止标示为正样本，其他的标示为负样本，那么正判负比负判正严重得多。
准确率不能够评价出正判负和负判正谁更严重。
召回率 := 正判正 / (正判正+正判负)
精确率 := 正判正 / (正判正+负判正)
召回率可以反应出正判负的严重程度，还有综合评价指标$F1$为召回率和精确率的加权调和平均：
令召回率R， 精确率P， 有$F1 = \frac{2PR}{P+R}$, $F1$值越大说明R和P的值都大。