西瓜书复习笔记06

- 间隔与支持向量
  - 什么是SVM：
    对于二分类问题，SVM就是在样本数据中计算一个最优超平面，然后把样本数据分割成不同的类别，并且能对新数据进行分类。
  - SVM和LR区别：
    - 相同点：
      都是分类算法
      都是监督学习
      都是判别式模型
      都是线性模型
    - 不同点：
      损失函数（目标函数）不同（LR：是基于概率理论和极大似然估计；SVM：是基于几何间隔最大化原理）。
      线性SVM是距离度量，需要normalization。
      SVM自带正则化，LR必须额外添加。
      SVM有核函数，LR一般不用核函数
  - 输出标签：
    二分类问题，标签是{-1, +1}
  - 划分超平面：
    将不同类别的养分分开的超平面。
    这个划分超平面受训练集局限性和噪点影响。
    这个划分超平面分类的结果是最鲁棒的，对未见样本泛化能力最强。
  - 划分超平面描述（线性方程）：

$$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b=0$$

    其中：
    w为法向量，决定了超平面的方向；
    b为位移项，决定了超平面与原点的距离。
  - 支持向量：
    支持平面上把两类划分开的超平面的向量点，即距离超平面最近的几个训练样本点。
  - 手推SVM：
    - 间隔：

$$r=\frac{\left|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b\right|}{\|\boldsymbol{w}\|}$$

其中：

$$\begin{array}{ll}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b\geqslant+1, & y_i=+1\\ \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b\leqslant-1, & y_i=-1\end{array}$$

$$\|\boldsymbol{w}\|=\sqrt{w_1^2+w_2^2+\cdots}$$

则有

$$\gamma=\frac{2}{\|\boldsymbol{w}\|}$$

- 最大间隔：

$$\begin{aligned}&\max_{\boldsymbol{w},b}\frac{2}{\|\boldsymbol{w}\|}\\ &\text{s.t. } y_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b\right)\geqslant 1,\quad i=1,2,\ldots,m\end{aligned}$$

等价于：

$$\begin{aligned}&\min_{\boldsymbol{w},b}\frac{1}{2}|\boldsymbol{w}\|\\ &\text{s.t. } y_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b\right)\geqslant 1,\quad i=1,2,\ldots,m\end{aligned}$$

这是SVM的基本型
我们希望对SVM的基础行求解得到模型：

$$f(x)=w^{\mathrm{T}}x+b$$

- 拉格朗日乘子法：
  - 凸二次规划：
    即优化(最小化或最大化)多个变量的二次函数，并服从于这些变量的线性约束。
  - 意义：
    - 在极值点，两函数的梯度在交点相切。

    - 梯度与等高线的切线是垂直的。
- 对偶问题：
  通过拉格朗日乘子法可将SVM基本型转为对偶问题
  目标函数：

$$L(w,b,\alpha)=\frac{1}{2}\|w\|^2+\sum_{i=1}^{m}\alpha_i\left(1-y_i\left(w^{\mathrm{T}}x_i+b\right)\right)$$

  其中：
  前面是正则项；
  对基本型的每条约束加上拉格朗日乘子$\alpha_i \geqslant 0$
  目标函数求w和b的偏导：

$$\begin{aligned}\boldsymbol{w}&=\sum_{i=1}^{m}\alpha_i y_i \boldsymbol{x}_i\\ 0&=\sum_{i=1}^{m}\alpha_i y_i\end{aligned}$$

  对偶问题：

$$\begin{gathered}\max_\alpha \textstyle\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^{\mathrm{T}}x_j\\ \text{s.t. } \textstyle\sum^{m}\alpha_i y_i=0\\ \alpha_i\geqslant 0,\quad i=1,2,\ldots,m\end{gathered}$$

- 最终模型：
  求解α，求出w和b可得：

$$\begin{aligned}f(x)&=w^{\mathrm{T}}x+b\\ &=\sum_{i=1}^{m}\alpha_i y_i x_i^{\mathrm{T}}x+b\end{aligned}$$

  求解过程必须满足KKT：

$$\left\{\begin{array}{l}\alpha_i\geqslant 0\\ y_i f\left(\boldsymbol{x}_i\right)-1\geqslant 0\\ \alpha_i\left(y_i f\left(\boldsymbol{x}_i\right)-1\right)=0\end{array}\right.$$

        其中：
        如果$\alpha_i = 0$，表示该样本不会在最终模型求和中出现，也就是这个样本不会影响到模型；
        如果$\alpha_i > 0$，表示该样本是支持向量，影响模型的形成；
        意义在于，训练完后，大多数训练样本不需要保留，最终模型仅与支持向量有关
    - 如何求解α：
      由于是二次规划问题，不能用梯度下降，常采用SMO。
        - 什么是SMO：
          在KKT约束情况下，SMO每次选择两个变量$\alpha_i\alpha_j$，并固定其他参数，求解获得更新后的$\alpha_i\alpha_j$，重复这个过程。总的来说$\alpha_i\alpha_j$差距越大越好。
- **核函数：**
    - 什么是核函数：
      对于线性不可分的问题，可以讲样本从原始空间映射到一个更高纬度的空间，使得样本在这个高维空间内线性可分。
    - 为什么要用核函数：
      训练样本是线性不可分的情况下（例如异或）可以采用核函数。
    - 除了核函数还有什么能处理噪点：
      核函数适合处理噪点过多的现象，软间隔适合处理噪点较少的现象。
    - 核函数会增加计算量吗：
      不会，核函数是在原始空间内计算内积，不必在高维空间计算内积。
    - 核函数的对偶问题：

$$
\begin{aligned}
& \max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \kappa\left(x_i, x_j\right) \\
& \text{s.t.} \sum_{i=1}^{m} \alpha_i y_i = 0 \\
& \qquad \alpha_i \geqslant 0, \quad i = 1, 2, \ldots, m
\end{aligned}
$$

    - 核函数选择：
      核函数的选择是支持向量机最大的变数，选择不好往往映射到一个不合适的空间，导致性能降低。
    - 常见的核函数：
      线性核
      多项式核
      高斯核
      拉普拉斯核
      Sigmoid核
- **软间隔：**

- 有了核函数为什么还要使用软间隔：
    - 线性可分，即能找到超平面，对于硬间隔支持向量机
    - 部分点不可分，总体近似可分，近似线性可分，对应软间隔支持向量机
    - 线性不可分，需要用到核函数
- 什么是软间隔：
  软间隔引入松弛变量，允许某些样本不满足硬间隔的约束条件。
- 损失函数：
  hinge损失：

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \max\left(0, 1 - y_i\left(w^{\mathrm{T}} x_i + b\right)\right)$$

$$\begin{aligned} \text{s.t.} \quad & y_i\left(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_i + b\right) \geqslant 1 - \xi_i \\ & \xi_i \geqslant 0, i = 1, 2, \ldots, m \end{aligned}$$

  其中：
  ξ_i为松弛变量，大于等于0，每个样本都对应一个松弛变量。
  C是一个常数，当C无限大，函数必须满足约束；当C是有限值时，允许一些样本不满足约束。
  求解：
  拉格朗日，但是与硬间隔的区别是KKT约束不同。
- 软间隔的模型与什么有关：
  仅与支持向量有关。
- 损失函数的正则化：
  L2范数倾向于w的分量取值尽可能的均匀，非零分量的个数尽量稠密；
  L1范数倾向于w分量尽可能的稀疏，分零分量尽可能的少。

- **多分类：**
  https://blog.csdn.net/csdn_lzw/article/details/80170178
- **SVM的回归问题：**