SVM--Support Vector Machine

七月在线 张雨石
2018年6月30日
http://blog.csdn.net/stdcoutzyx

SVM

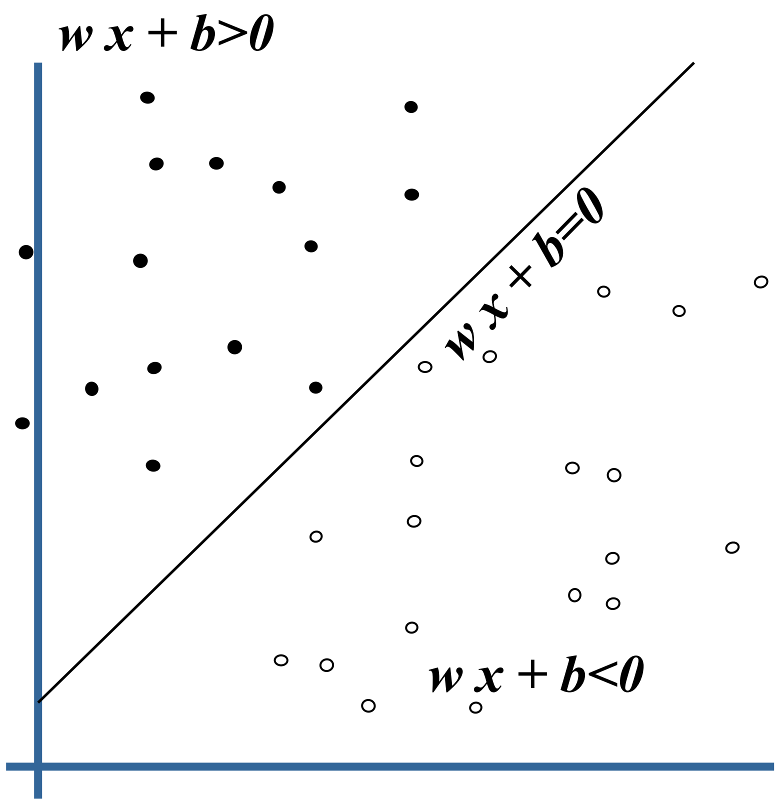
- SVM模型
  - 线性可分分类器
  - 线性不可分分类器
  - SMO求解
- SVM实战文本分类

SVM模型

- 函数间隔与几何间隔
- 最优间隔分类器
- 拉格朗日求解
- 最优间隔分类器求解
- SMO算法
- 核技法
- 软间隔分类器
- 合页损失函数
- 多分类

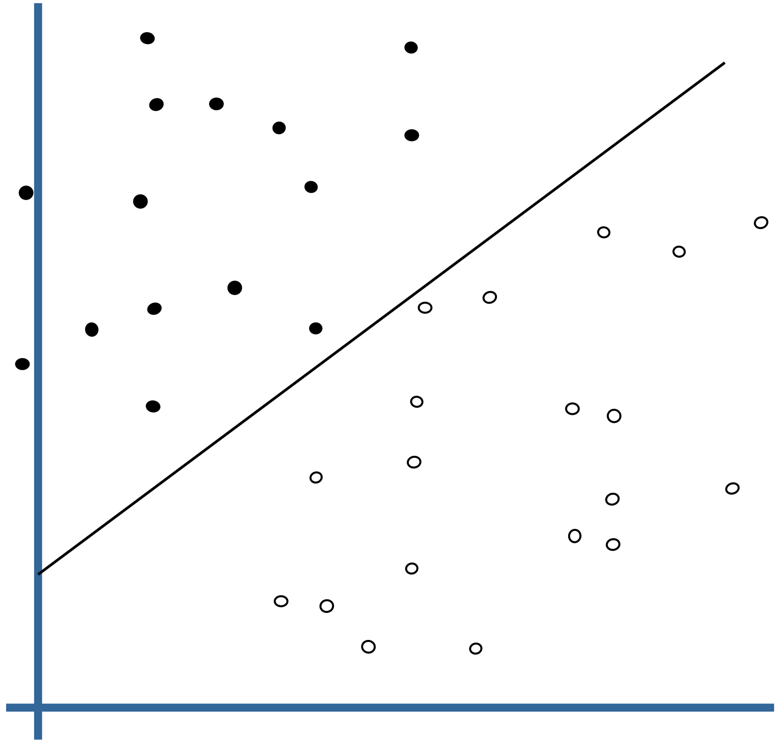
函数间隔与几何间隔

- • 表示 +1类
- ○ 表示 -1 类



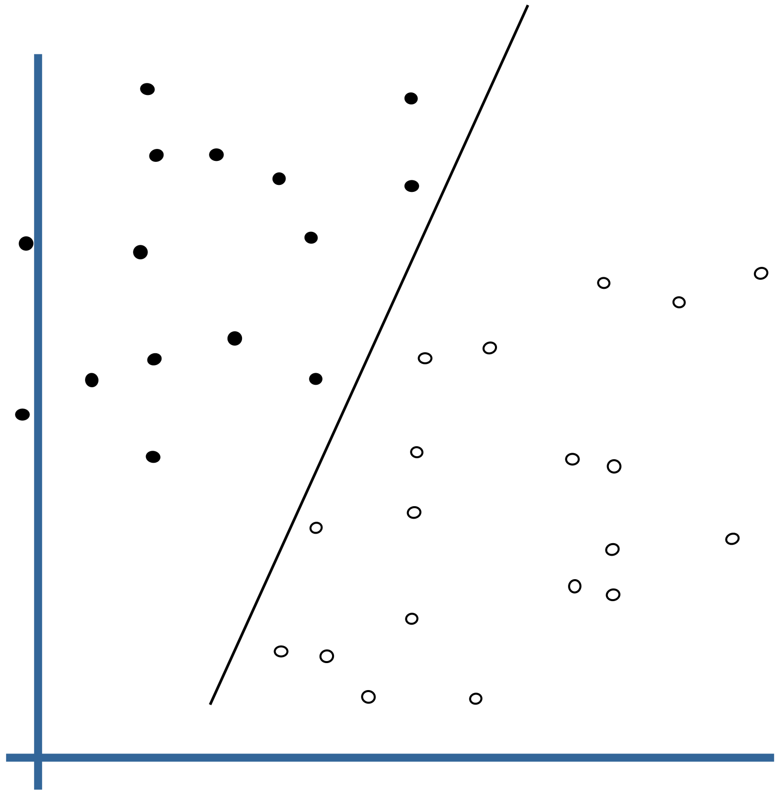
函数间隔与几何间隔

- •表示 +1类
- ◦表示 -1 类



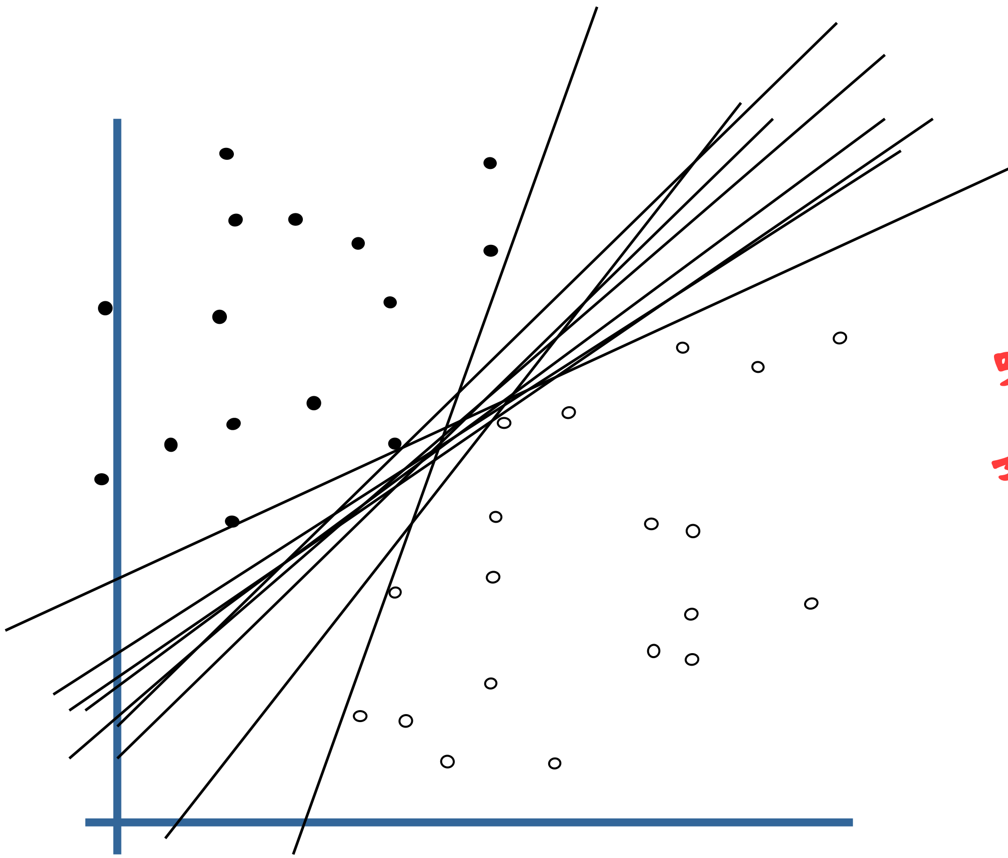
函数间隔与几何间隔

- • 表示 +1类
- ○ 表示 -1 类



函数间隔与几何间隔

- 表示+1
- ◦ 表示 -1



哪一条线是最好的？

函数间隔与几何间隔

- 公式化问题
  - 分类模型

$$g(z)=\begin{cases}-1 & z<0\\ 1 & z\geq 0\end{cases}$$

$$h_{w,\mathrm{b}}(x)=g(w^Tx+b)$$

  - 函数间隔

$$\hat{\ell}^{(\mathrm{i})}=y^{(i)}(w^Tx^{(i)}+b)$$

$$\hat{\ell}=\min_i \hat{\ell}^{(\mathrm{i})}$$

函数间隔与几何间隔

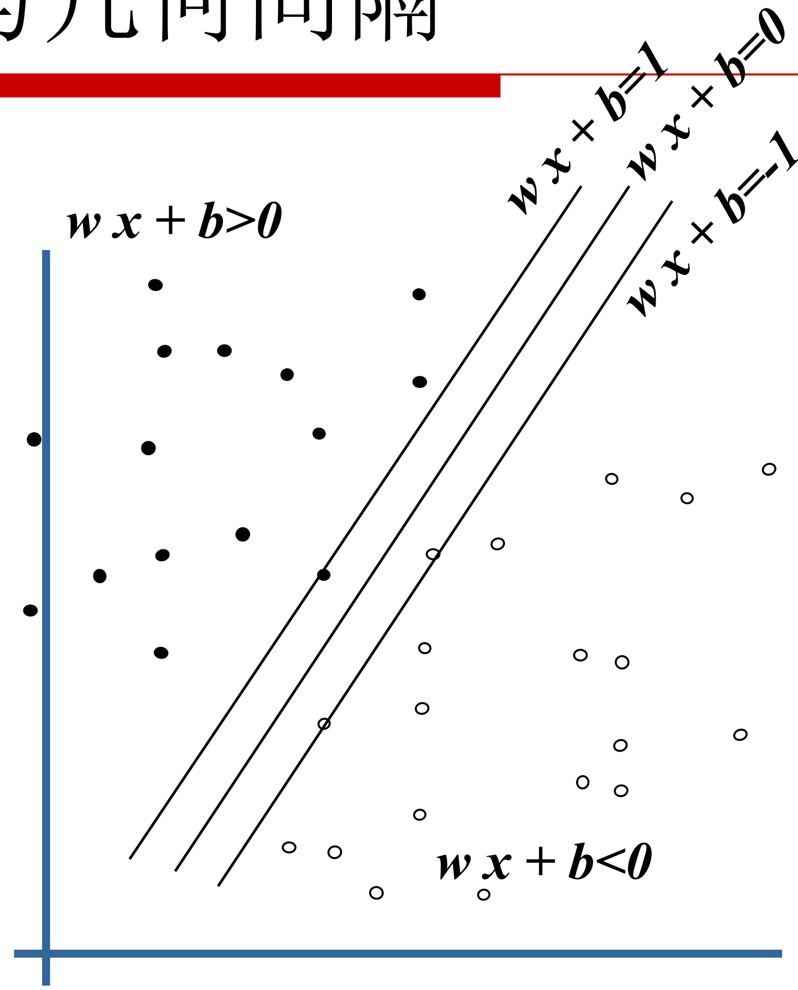
- 只要成倍的增大w和b函数间隔就能无限增大
- 几何间隔
  - 限定w

$$\max_{w,b} \ell \qquad s.t.\ y^{(i)}\left(w^T x^{(i)} + b\right) \geq \ell \quad \text{且} \ ||\mathrm{w}|| = 1$$

  - 即在||w||=1 条件下函数间隔最小值

函数间隔与几何间隔

- •表示 +1类
- ○表示 -1 类



最优间隔分类器

- 初始函数表达

$$
\begin{aligned}
&max_{\gamma,w,b} \quad \gamma \\
&\text{s.t.} \quad y^{(i)}\left(w^T x^{(i)} + b\right) \geq \gamma,\ \ i = 1,2,\dots,m \\
&||\mathrm{w}|| = 1
\end{aligned}
$$

  - 非凸性约束，容易达到局部最优

最优间隔分类器

简化-1

$$\max_{\gamma,w,b} \quad \frac{\gamma}{||w||}$$

$$\cancel{\mathrm{s.t.} \quad y^{(i)}\left(\frac{w^T}{||w||}x^{(i)}+\frac{b}{||w||}\right)\geq\frac{\gamma}{||w||},\ \ i=1,2,\ldots,m}$$

$$\mathrm{s.t.} \quad y^{(i)}\left(w^T x^{(i)}+b\right)\geq\gamma,\ \ i=1,2,\ldots,m$$

最优间隔分类器

- 简化-2
  - 调整w，b，可以方便的将r变为1，所以索性直接变为1

$$\cancel{max_{\gamma,w,b} \quad \frac{1}{||w||}}$$

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

$$\text{s.t.} \quad y^{(i)}\left(w^T x^{(i)} + b\right) \geq 1, \;\; i = 1,2,\ldots,m$$

最优间隔分类器

- 最终问题

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

$$\text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq 1, \;\; i = 1,2,\ldots,m$$

- 线性约束下优化二次函数
  - 有数学方法可以解决这个问题
  - 引入对偶函数

拉格朗日函数

- 回顾高数课程 $\min_w \quad f(w)$

$$s.t.\ h_i(w) = 0, i = 1,2, \ldots, l$$

- 构造拉格朗日函数

$$L(w, \beta) = f(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

- 求解

$$\frac{\partial L}{\partial w_i} = 0; \frac{\partial L}{\partial \beta_i} = 0$$

广义拉格朗日函数

- 有不等式约束的时候

$$
\begin{aligned}
&min_w \quad f(w) \\
&\qquad s.t. \ h_i(w) = 0, i = 1,2,\dots,l \\
&\qquad\qquad g_i(w) \le 0, i = 1,2,\dots,k
\end{aligned}
$$

- 构建拉格朗日方程

$$L(w,\alpha,\beta) = f(w) + \sum_{i=1}^{l} \beta_i h_i(w) + \sum_{i=1}^{k} \alpha_i g_i(w)$$

广义拉格朗日函数

- 极小极大

$$\theta_p(w) = \max_{\alpha,\beta:\alpha_i>0} f(w) + \sum_{i=1}^{l} \beta_i h_i(w) + \sum_{i=1}^{k} \alpha_i g_i(w)$$

$$\theta_p(w) = \begin{cases} f(w) & \text{所有约束都满足} \\ \infty & \text{否则} \end{cases}$$

$$\min_{w} \theta_p(w) = \min_{w} \max_{\alpha,\beta:\alpha_i>0} L(w,\alpha,\beta)$$

广义拉格朗日函数

- 极小极大的最优解

$$p^* = min_w \theta_p(w)$$

广义拉格朗日函数

- 对偶问题——极大极小

$$\theta_D(\alpha,\beta) = \min_w L(w,\alpha,\beta)$$

$$\max_{\alpha,\beta:\alpha_i>0} \theta_D(\alpha,\beta) = \max_{\alpha,\beta:\alpha_i>0} \min_w L(w,\alpha,\beta)$$

$$d^* = max_{\alpha,\beta:\alpha_i>0}\theta_D(\alpha,\beta)$$

广义拉格朗日函数

- 对偶问题与原始问题的等价性

$$d^* \leq p^*$$

- 约束不等式g都是凸函数
  - 线性函数都是凸函数
- 约束等式h都是仿射函数
  - 仿射和线性等价，除了允许截距b
- 不等式严格执行
  - 必有g不等式是小于0的

广义拉格朗日函数

- 对偶问题与原始问题等价性
  - 在上述假设下，只要满足KKT条件，最优解相等
    - 即存在W* a* b*
    - W*是原始问题的解
    - A* b*是对偶问题的解

$$p^* = d^* = L(w^*, \alpha^*, \beta^*)$$

广义拉格朗日函数

- KKT条件

$$L(w,\alpha,\beta)=f(w)+\sum_{i=1}^{l}\beta_i h_i(w)+\sum_{i=1}^{k}\alpha_i g_i(w)$$

$$\frac{\partial}{\partial w_i}L(w^*,\alpha^*,\beta^*)=0, i=1,2,\ldots,n$$

$$\frac{\partial}{\partial \beta_i}L(w^*,\alpha^*,\beta^*)=0, i=1,2,\ldots,l$$

$$\alpha_i^* g_i(w^*)=0, i=1,2,\ldots,k$$

$$g_i(w^*)\leq 0, i=1,2,\ldots,k$$

$$\alpha^*\geq 0, i=1,2,\ldots,k$$

最优间隔分类器求解

- 使用拉格朗日方程

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

$$\text{s.t.} \quad g_i(w) = -y^{(i)}\left(w^T x^{(i)} + b\right) + 1 \le 0, \quad i = 1,2,\dots,m$$

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

最优间隔分类器求解

- 解决对偶问题

$$\max_{\alpha,\beta:\alpha_i>0} \theta_D(\alpha,\beta) = \max_{\alpha,\beta:\alpha_i>0} \min_{w} L(w,\alpha,\beta)$$

最优间隔分类器求解

- 先固定alpha、beta，对w和b求导

$$\nabla_w L(\mathrm{w}, \mathrm{b}, \alpha) = w - \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} = 0 \Longrightarrow w = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}$$

$$\frac{\partial}{\partial b} L(\mathrm{w}, \mathrm{b}, \alpha) = \sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

$$\mathrm{L}(\mathrm{w},\mathrm{b},\alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{m} \alpha_i [y^{(i)}(w^T x^{(i)} + b) - 1]$$

$$= \frac{1}{2} w^T w - \sum_{i=1}^{m} \alpha_i y^{(i)} w^T x^{(i)} - \sum_{i=1}^{m} \alpha_i y^{(i)} b + \sum_{i=1}^{m} \alpha_i$$

$$= \frac{1}{2} w^T \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} - \sum_{i=1}^{m} \alpha_i y^{(i)} w^T x^{(i)} + \sum_{i=1}^{m} \alpha_i$$

$$= \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \alpha_i y^{(i)} w^T x^{(i)}$$

$$= \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \alpha_i y^{(i)} \left(\sum_{j=1}^{m} \alpha_j y^{(j)} (x^{(j)})^T \right) x^{(i)}$$

最优间隔分类器求解

- 问题变换

$$L(w,b,\alpha) = \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y^{(i)}y^{(j)}(x^{(j)})^T x^{(i)}$$

$$\max_{\alpha} \quad \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y^{(i)}y^{(j)}\langle x^{(j)}, x^{(i)}\rangle$$

$$\text{s.t.} \quad \alpha_i \geq 0, \mathrm{i} = 1,2,\ldots,\mathrm{m}$$

$$\sum_{i=1}^{m}\alpha_i y^{(i)} = 0$$

最优间隔分类器求解

- 每一个非0的alpha预示着这是一个支持向量
- 分类模型为

$$f(x) = w^T x + b = \sum a_i y^{(i)} x^{(i)^T} x + b$$

- 新数据的分类需要和所有支持向量做内积
- 解决这个问题需要训练集内所有的pair计算内积

SMO算法

- 坐标上升法

$$\max_{\alpha} W(\alpha_1, \alpha_2, \dots, \alpha_n)$$

Loop Until Convergence: {

For i=1,2…,m{

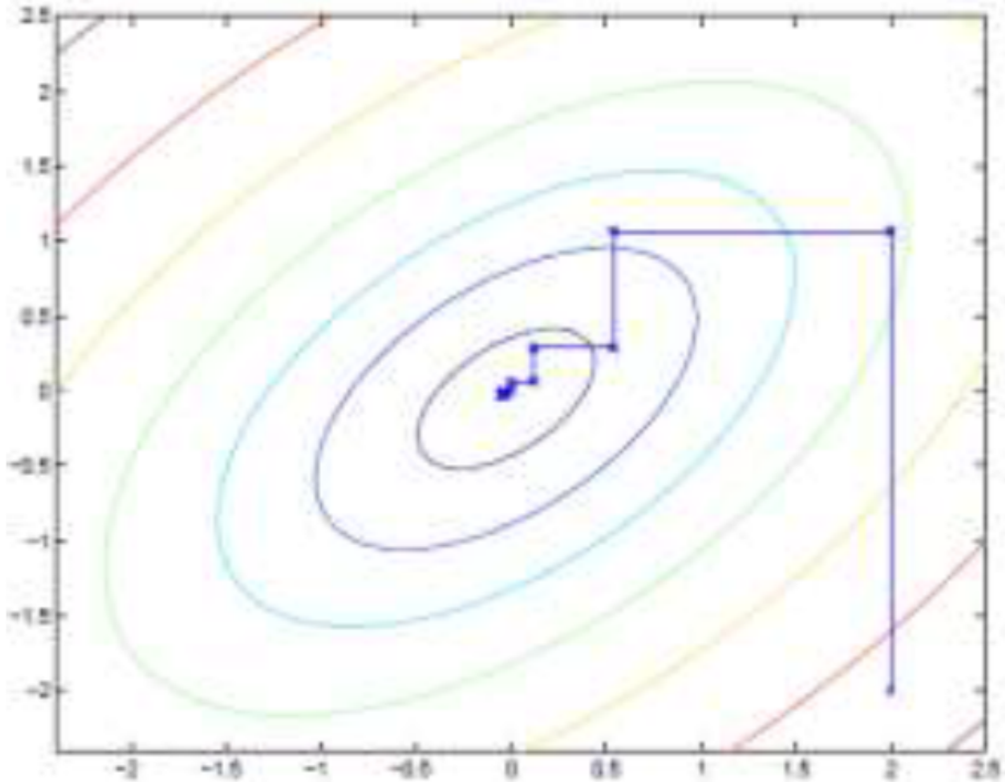
$$\alpha_i := argmax_{\hat{\alpha}_i} W(\alpha_1, \dots, \alpha_{i-1}, \hat{\alpha}_i, \alpha_{i+1}, \dots, \alpha_n)$$

}

}

SMO算法

- 二维坐标上升



SMO算法

- SMO（Sequential Minimal Optimization）
- 求解的问题有一个约束
  - 每次选择两个变量来进行优化

$$\max_{\alpha} \quad \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y^{(i)}y^{(j)}\langle x^{(j)}, x^{(i)}\rangle$$

$$\text{s.t.} \quad \alpha_i \geq 0, \mathrm{i} = 1,2,\dots,\mathrm{m}$$

$$\sum_{i=1}^{m}\alpha_i y^{(i)} = 0$$

SMO算法

- 两个变量其实等价一个变量

$$\alpha_2 = y^{(2)}(-\sum_{i=3}^{m}\alpha_i y^{(i)} - \alpha_1 y^{(1)})$$

- 算法流程

Repeat till Convergence{

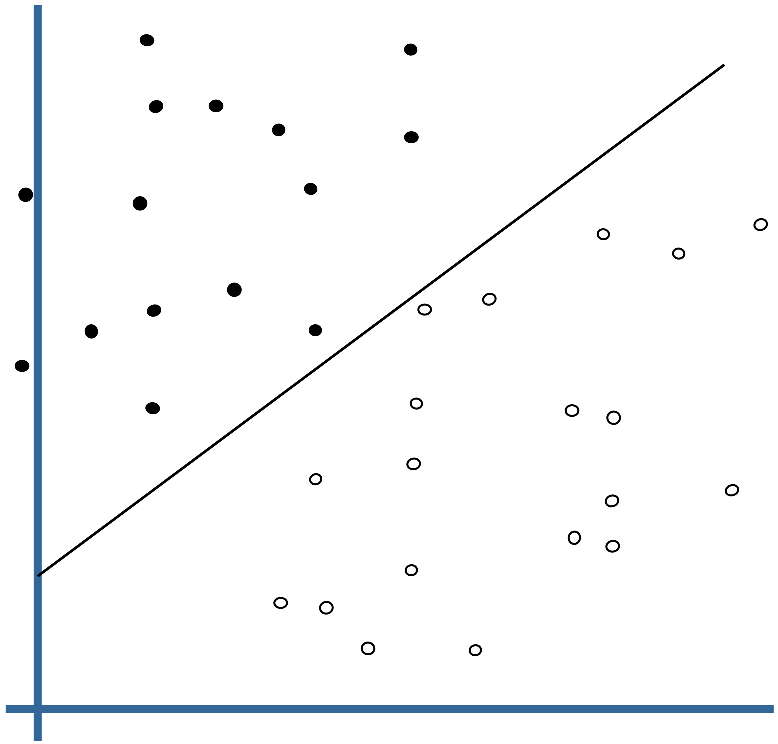
Select two parameter $\alpha_i \alpha_j (i \neq j)$

Optimize W(α) with respect to $\alpha_i \alpha_j$, holding other parameters fixed

}

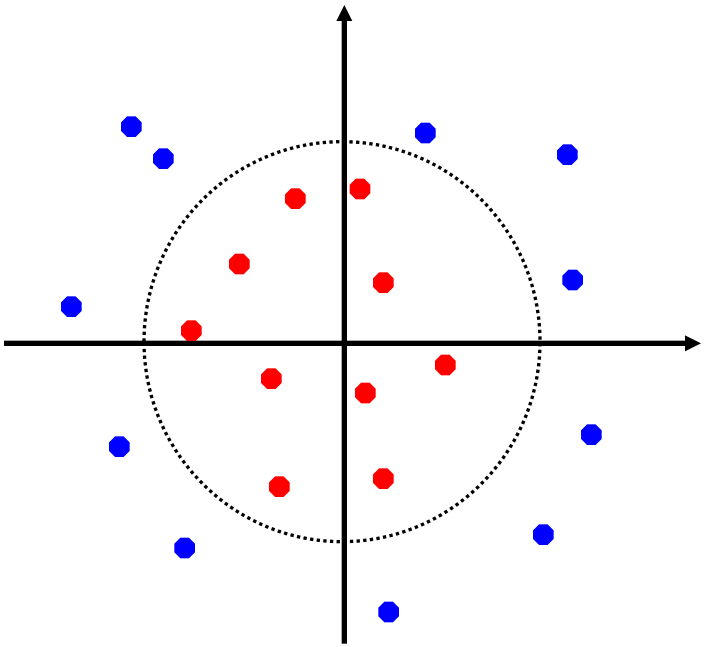
核技法

- 回想原始问题——线性可分



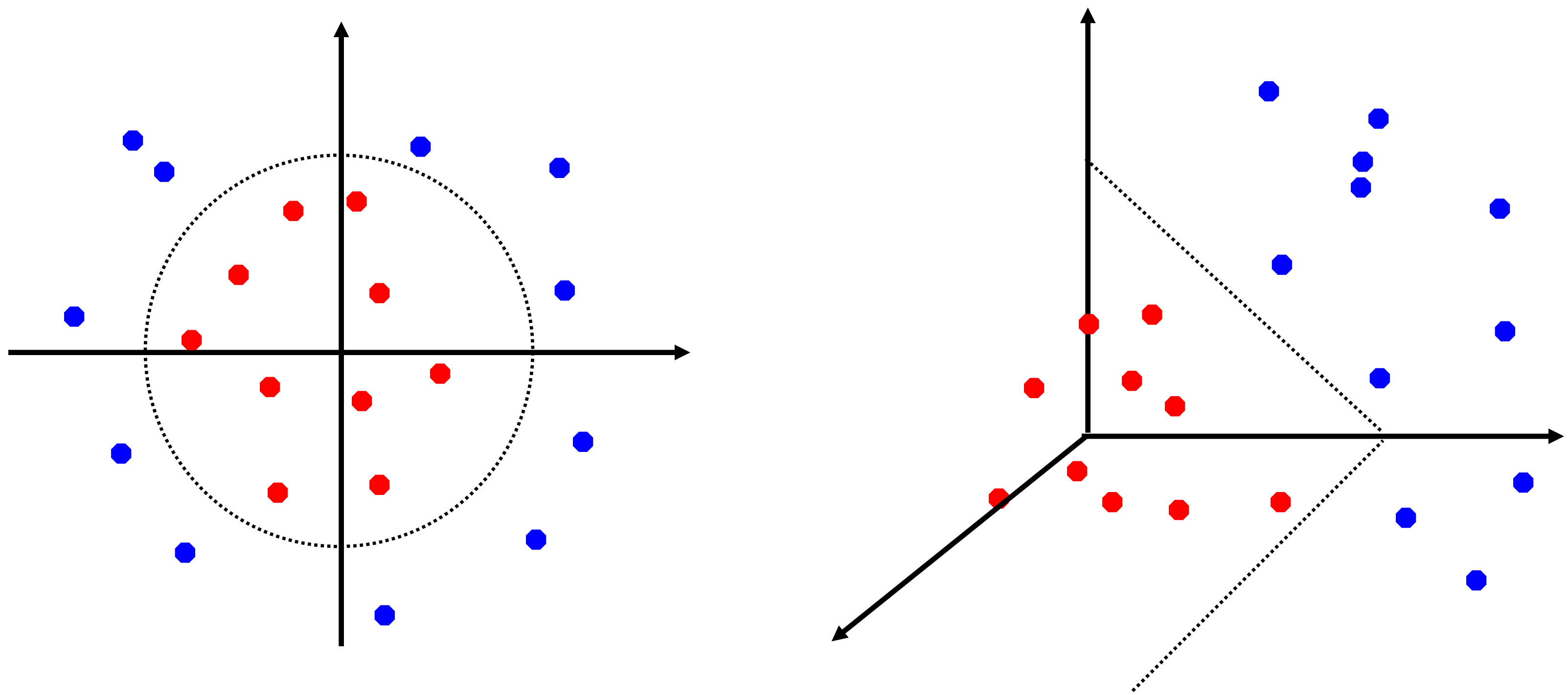
核技法

- 线性不可分问题



核技法

- 线性不可分问题——空间变换



核技法

- 空间变换公式化

$$\max_{\alpha} \quad \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y^{(i)}y^{(j)}\langle x^{(j)}, x^{(i)}\rangle$$

$$\text{s.t.} \quad \alpha_i \geq 0, \mathrm{i} = 1,2,\dots,\mathrm{m}$$

$$\sum_{i=1}^{m}\alpha_i y^{(i)} = 0$$

$$\mathrm{K(X,Z)} = \langle\phi(X), \phi(Z)\rangle$$

核技法

- 核函数对应一种映射函数
- 为什么使用核函数

$$K(X,Z) = \langle \phi(X), \phi(Z) \rangle$$

- 经过映射后的向量维度可能过高，导致向量内积计算量过大

核技法

- 核函数例子-1

$$K(X,Z) = (X^T Z)^2 = \left(\sum_{i=1}^{n} x_i z_i\right)\left(\sum_{j=1}^{n} x_j z_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{n}(x_i x_j)(z_i z_j)$$

- 对应映射函数

$$\phi(x) = [x_1 x_1 \quad x_1 x_2 \quad \cdots \quad x_n x_{n-1} \quad x_n x_n]^T$$

核技法

- 核函数例子-2

$$K(X,Z) = (X^T Z + c)^2 = \sum_{i,j=1}^{n} (x_i x_j)(z_i z_j) + \sum_{i=1}^{n} \sqrt{2c}\, x_i \sqrt{2c}\, z_i + c^2$$

- 对应的映射函数

$$\phi(x) = [x_1 x_1 \quad \dots \quad x_n x_n \quad \sqrt{2c}\, x_1 \quad \dots \quad \sqrt{2c}\, x_n \quad c]^T$$

核技法

- 核函数作用
  - 例子1和例子2都体现出了核函数降低计算量
  - 另一个层面，如果x1和x2在对应维度空间位置接近，那么内积很大。所以核函数K是接近程度的度量函数
    - 高斯核——对应无限维

$$K(x,z) = \exp\left(-\frac{||x-z||^2}{2\sigma^2}\right)$$

核技法

- 什么样的核是合法的？
  - 定义一个核矩阵K $K_{ij} = K(x^{(i)}, x^{(j)})$
  - Mercer定理
    - K是合法的核的充分必要条件是对于一个有限的数据集，对应的核矩阵都是对称半正定矩阵

核技法

- 常用核函数
  - 多项式

$$K(x^{(i)}, x^{(j)}) = (1 + {x^{(i)}}^T x^{(j)})^p$$

  - 高斯

$$K(x^{(i)}, x^{(j)}) = \exp(-\frac{||x^{(i)} - x^{(j)}||^2}{2\sigma^2})$$

  - sigmoid

$$K(x^{(i)}, x^{(j)}) = \tanh(\beta_0 x^{(i)} {x^{(j)}}^T + \beta_0)$$

核技法

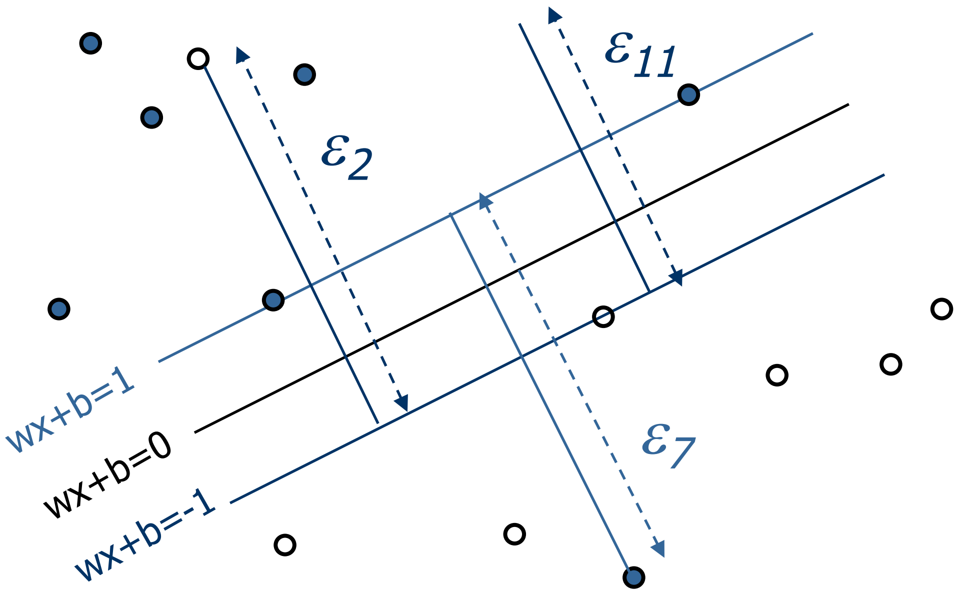
- SVM确定一个超平面来进行分类
- 如果当前空间不是线性可分，映射到高维空间
  - SVM不直接进行映射，而是利用核函数
- 核函数应用在向量内积上

软间隔分类器

- 当数据线性不可分时
  - 映射到高维空间
  - 高维映射并不能保证数据线性可分，只能说有更大概率线性可分

软间隔分类器

- 高维空间仍然线性不可分时?



软间隔分类器

- 允许有数据点拥有小于1的几何间隔
- 但要受到惩罚

$$min_{w,b} \quad \frac{1}{2}||w||^2 + \mathrm{C}\sum_{i=1}^{m}\varepsilon_i$$

$$\mathrm{s.t.} \quad y^{(i)}\left(w^T x^{(i)} + b\right) \geq 1 - \varepsilon_i, \ \ i = 1,2,\ldots,m$$

$$\varepsilon_i \geq 0, i = 1,2,\ldots,m$$

软间隔分类器

- 拉格朗日方程

$$L(w, b, \varepsilon, \alpha, r) = \frac{1}{2}||w||^2 + C\sum_{i=1}^{m}\varepsilon_i$$

$$-\sum_{i=1}^{m}\alpha_i\left[y^{(i)}\left(w^T x^{(i)} + b\right) - 1 + \varepsilon_i\right] - \sum_{i=1}^{m} r_i\varepsilon_i$$

软间隔分类器

- 新的对偶问题
  - 和原来相比，只对alpha做了更多约束

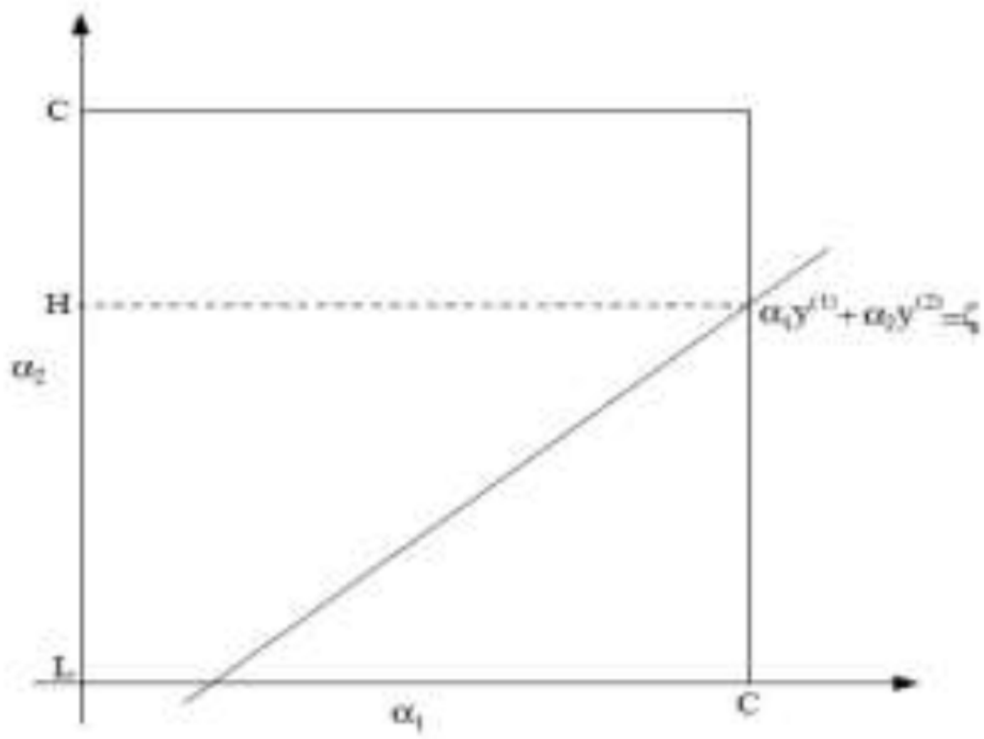
$$\max_\alpha \quad W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} \langle x^{(j)}, x^{(i)} \rangle$$

$$\text{s.t.} \quad 0 \le \alpha_i \le \mathrm{C}, \mathrm{i} = 1,2, \ldots, \mathrm{m}$$

$$\sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

软间隔分类器

- 新约束下SMO算法



SVM性质

- 数学特性
  - 凸优化问题，保证会有全局最优解
- 模型特性
  - 可以处理高维数据
  - 软间隔降低过拟合
  - 求解完成后只有少数数据起作用
  - 灵活的选择核函数

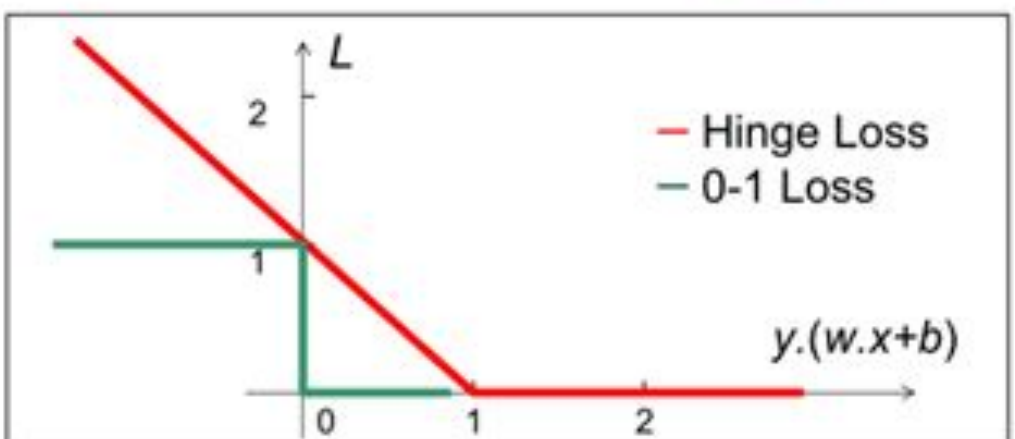
合页损失函数（hinge loss）

- Svm的另一种理解

$$[z]_+ = \begin{cases} z \; if \; z > 0 \\ 0 \end{cases}$$

$$\text{Loss} = \sum_{i=1}^{N} [1 - y^{(i)}(w^T x^{(i)} + b)]_+ + \lambda ||w||^2$$

$\min_{\{w,b\}} L(w,b) + \lambda \, ||w||^2$ $\lambda = 0.5/C$



多分类

- 现有的SVM仅支持二类分类
- 多(N)分类解决
  - 一对多
    - N个分类器
  - 一对一
    - N(N-1)/2个分类器
  - 层次支持向量机
    - LogN个分类器

实战文本分类

- 特征工程+分类
- 深度学习-端到端

实战文本分类

- 项目-百度关键词分类比赛（2013）
  - 问题：百度搜索串分类，算法基于hadoop实现
  - 数据量：
    - 训练数据：1000w条，33类
    - 测试数据：100w条
  - 结果：
    - 97.65%（三等奖），一等奖98.65%

实战文本分类

- 问题处理流程
  - 文本分词
  - 特征筛选
    - 去除停用词
    - 特征重要程度计算
  - 分词结果向量化
  - LibSVM训练模型
  - 预测

实战文本分类

- 特征选择
  - TF-IDF
  - 词频
  - 文档频率
  - 互信息
  - 信息增益
  - 卡方分布

实战文本分类

- 分组数目与分类性能的权衡（0.05%-0.15%）
- 细粒度分词（0.8%左右）
  - 张三/说的/确实/在理
  - 张三/三/说的/的确/确实/实在/在理
- 向量化权重（0.02%）
- svm参数（0.2%）
  - -s 4 （MCSVM_CS，Multi-class SVM by Crammer and Singer）
- 停用词（0.04%）

实战文本分类

- Top1-解决方案-特征工程
  - 4-gram字符组合（180w）
    - 比如对于“生日蛋糕”这样的词语，提取“生日蛋糕”，“生日蛋”，“日蛋糕”，“生日”，“日蛋”，“蛋糕”，以及“生”，“日”，“蛋”，“糕”这些词语。
  - N-gram词语组合（570w）
    - 比如对于词语“天津 新开河街 房价”这样的短语我们会提取“天津新开河街”，“天津房价”，“新开河街房价”，“天津”，“新开河街”，“房价”这样的组合

实战文本分类——hadoop实现

- 分词-IKAnalyzer
- SVM
  - LibLinear
  - Hdfs文件读取
  - 一对一训练or分组训练
  - 训练map-reduce伪代码

```
class ClassifyMapper
    method setup
        train_data = svm.read(group_i, group_j)
        model = svm.train(train_data)
    method map(test_instance)
        label = svm.predict(model, test_instance)
```

实战文本分类

- Notebook

不平衡文本分类

- 数据层面
  - 重采样——上采样、下采样、SMOTE
  - 训练集划分———大类划分子集
- 传统算法层面
  - Random forest
  - 少数类加权
  - 多层分类
  - 规则集成

Thanks!

Q&A