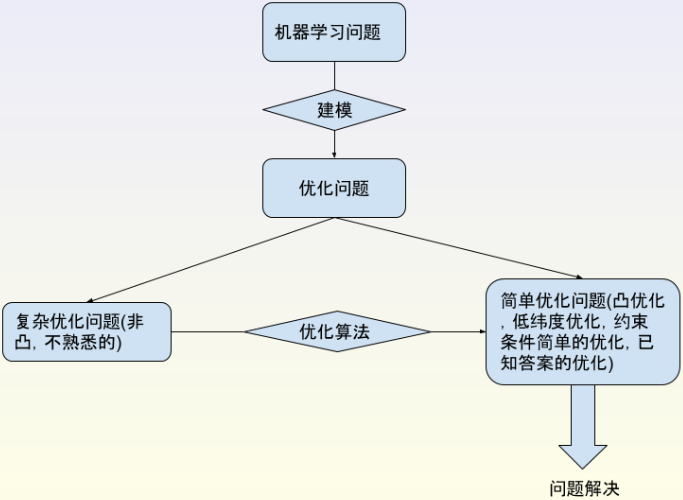
矩阵论与凸优化第 6 课凸优化在机器学习中的应用

管老师

七月在线

June, 2016

机器学习与优化


机器学习问题
建模
优化问题
复杂优化问题(非凸，不熟悉的)
优化算法
简单优化问题(凸优化，低纬度优化，约束条件简单的优化，已知答案的优化)
问题解决

主要内容

凸优化应用举例

本节课介绍凸优化在机器学习以及图像处理的两个应用，准备过程中参考了 YouTube 上的如下资源.

- 支持向量机 (SVM)
  - MIT 公开课:人工智能第 16 课 SVM
- 压缩感知与图像处理
  - 陶哲轩演讲:Compressed Sensing

支持向量机 (SVM)

支持向量机是针对分类问题的一种线性分类算法

分类问题

通过样本的特征数据，对样本进行分类的问题。

- 医疗诊断
- 网络信息安全
- 图像识别
- ...

支持向量机 (SVM)

每一个样本都可以看成是特征空间 $\mathcal{R}^n$ 中的一个点，一个分类算法则将空间分为若干部分，每部分都对应一个分类。

线性分类器

线性分类器泛指使用超平面来对特征空间进行分割的分类算法。

对于一些线性可分的例子来说，可以达到分类目的的线性分类器并不唯一，那么如何选择一个我们认为是“最好”的呢?

线性分类器的数学形式

空间 $\mathbb{R}^n$ 中有可分的两个点集 C,D. 线性分类器构造一个超平面 $a^Tx = b$ 对空间进行分割，使得 C 和 D 分别位于这个超平面的两侧.

$$a^Tp > b, \quad a^Tq < b$$

这等价于存在一个正数 t 使得 $a^Tp - b \geq t, \quad a^Tq - b \leq -t$, 而这个正数 t 一定程度上描述了两个集合被分开的程度. 所以一个比较好的想法就是希望当 a 是单位向量的时候，t 越大越好. 这样我们就得到了一个优化问题.

支持向量机 (SVM)

SVM 优化问题

$$
\begin{aligned}
\text{最小化:} \quad & -t \\
\text{不等条件 1:} \quad & -a^T p_i + b \le -t \\
\text{不等条件 2:} \quad & a^T q_i - b \le -t \\
\text{不等条件 3:} \quad & -t \le 0 \\
\text{等式条件:} \quad & |a|^2 = 1
\end{aligned}
$$

但是这里面 $|a|^2 = 1$ 并非凸条件，所以我们做一个变量代换 $w = a/t$，那么我们最大化 t 就等于最小化 $|w|^2$. 于是我们得到了下面的等价优化问题。

支持向量机 (SVM)

SVM 凸优化问题

$$\begin{aligned} \text{最小化:} \quad & \frac{1}{2}|w|^2 \\ \text{不等条件 1:} \quad & -w^T p_i + b \leq -1 \\ \text{不等条件 2:} \quad & w^T q_i - b \leq -1 \end{aligned}$$

为了方便表述, 将 p_i, q_i 统一定义为 x_i, 并定义变量 y 使得如果 $x_i \in C$ 则 $y_i = -1$, 否则 $y_i = 1$

SVM 凸优化问题

$$
\begin{aligned}
\text{最小化:} \quad & \frac{1}{2}|w|^2 \\
\text{不等条件 1:} \quad & y_i(w^T x_i - b) + 1 \leq 0
\end{aligned}
$$

SVM 凸优化对偶问题: 拉格朗日量

$$L(w,b,\lambda) = \frac{1}{2}|w|^2 + \sum_i \lambda_i (y_i(w^T x_i - b) + 1)$$

对偶函数为拉格朗日量在 w, b 上取到的极小值, 所以我们计算 L 的梯度:

$$\frac{\partial}{\partial w} L = w + \sum_i \lambda_i y_i x_i$$

$$\frac{\partial}{\partial b} L = \sum_i \lambda_i y_i$$

于是我们得到对偶函数

SVM 凸优化对偶问题: 对偶函数

$$
\begin{aligned}
g(\lambda) = &\frac{1}{2}(\sum_i \lambda_i y_i x_i)^T(\sum_i \lambda_i y_i x_i) \\
&+ \sum_i \lambda_i(y_i((-\sum_i \lambda_i y_i x_i)^T x_i - b) + 1) \\
&= \sum_i \lambda_i - \frac{1}{2}\sum_i \sum_j \lambda_i \lambda_j y_i y_j x_i^T x_j
\end{aligned}
$$

支持向量机 (SVM)

进一步简化, 定义:

$$
\begin{aligned}
y &= [y_1, y_2, \cdots, y_n]^T \\
Y &= \mathrm{diag}(y_1, y_2, \cdots, y_n) \\
X &= [x_1, x_2, \cdots, x_n] \\
E &= [1, 1, \cdots, 1]^T \\
K_0 &= Y^T X^T X Y
\end{aligned}
$$

于是

SVM 凸优化对偶问题: 对偶函数

$$g(\lambda) = \lambda^T E - \frac{1}{2}\lambda^T K_0 \lambda$$

支持向量机 (SVM)

SVM 凸优化对偶问题

$$
\begin{aligned}
\text{最大化:}\quad & \lambda^T E - \frac{1}{2}\lambda^T K_0 \lambda \\
\text{不等条件:}\quad & \lambda_i \geq 0 \\
\text{定义域:}\quad & \lambda^T y = 0
\end{aligned}
$$

K_0 集中包含了 x_i 的全部信息. 不同的分类问题对应着不同的 K_0. 对于非线性 SVM 问题，所谓 kernel 技巧，就是针对 K_0 进行变换 (此处不再详述).

支持向量机 (SVM)

考虑 C 与 D 的凸包 $\overline{C}$ 和 $\overline{D}$

SVM 凸集合思路

若 $p \in \overline{C}, q \in \overline{D}$, 满足 $d(p,q) = d(\overline{C}, \overline{D})$ 则，向量 $\vec{pq}$ 即为所求支撑向量。

支持向量机 (SVM)

定义:

$$P = [p_1, \cdots, p_{n_c}]$$
$$Q = [q_1, \cdots, q_{n_D}]$$

于是:$X^T = [P^T, Q^T]$, 而对于 $\lambda^T P \in \overline{C}$ 与 $\mu^T Q \in \overline{D}$, 这两点之间的距离为

$$(\lambda^T P - \mu^T Q)^T (\lambda^T P - \mu^T Q) = [\lambda^T, \mu^T] K_0 [\lambda^T, \mu^T]^T$$

支持向量机 (SVM)

SVM 凸优化问题 (凸集合思路)

$$
\begin{aligned}
\text{最小化:} \quad & [\lambda^T, \mu^T] K_0 [\lambda^T, \mu^T]^T \\
\text{不等条件:} \quad & \lambda \geq 0 \\
\text{不等条件:} \quad & \mu \geq 0 \\
\text{等式条件:} \quad & \lambda^T E = 1 \\
\text{等式条件:} \quad & \mu^T E = 1
\end{aligned}
$$

再一次我们看到 K_0 集中包含了 x_i 的全部信息. 不同的分类问题对应着不同的 K_0.

思考题: 请证明凸集合思路的 SVM 优化问题与前面的 SVM 凸优化对偶问题等价

压缩感知与图像处理

压缩感知 (compressive sensing) 是一个关于信号还原的理论, 其基本问题是如下的一个线性方程

线性代数问题

A 是一个 $m \times n$ 的矩阵, b 是一个 m 维向量, 求解方程 $Ax = b$.

其中 b 是接收到的信号, x 是待还原的信号。

压缩感知与图像处理

线性方程经典结论

i 如果 $m = n$, 而且 A 是可逆的，那么 $x = A^{-1}y$

ii 如果 $m > n$，那么我们使用最小二乘进行估计 $x \sim (A^T A)^{-1} A^T b$

iii 如果 $m < n$, 则通常解不唯一，$x = x_0 + l$, $l \in \mathsf{Null}(A)$

压缩感知关心的正是第 iii 种情况！

压缩感知有可能实现吗?

一般情况下不可能

如果 b 的维数小于 x 的维数, 一般而言，方程 $Ax = b$ 有无穷多解。所以不可能通过 b 来恢复 x.

当 x 具有稀梳性的时候，可能实现压缩感知

如果已知原信号 x 中只有 S 个元素不为零，那么称 x 为 $S - Sparse$. 在这种情况下，$Ax = b$ 有可能具有唯一一个满足 x 的稀梳性的解.

压缩感知的理论保障

如果 A 的任意 $2S$ 列都线性无关，那么 $Ax = b$ 的 S-sparse 解具有唯一性.

证明

压缩感知与图像处理

压缩感知的适用范围

- 接收信号 b 维数小于原信号 x 的维数
- 原信号 x 具有稀梳性 (Sparsity)

- 医学图像 (核磁共振，CT 成像)
- 石油探测成像
- 单像素相机
- $\cdots$

压缩感知的实现

在 $Ax = b$ 全部解中寻找满足稀梳性 S-sparse 的解。

稀梳性是 x 非零元素的个数，于是得到优化问题.

压缩感知与图像处理

压缩感知优化问题

$$\begin{aligned}\text{最小化:}\quad & |x|_0 \\ \text{等式条件:}\quad & Ax-b=0\end{aligned}$$

优化函数 $|x|_0$ 并非凸函数，很难求解。

压缩感知优化问题

$$\begin{aligned}\text{最小化:}\quad & |x|_2 \\ \text{等式条件:}\quad & Ax-b=0\end{aligned}$$

优化函数 $|x|_2$ 很容易求解，但是不能得到原问题的解。

压缩感知与图像处理

压缩感知优化问题

$$\begin{aligned}\text{最小化:}\quad & |x|_1\\ \text{等式条件:}\quad & Ax-b=0\end{aligned}$$

优化函数 $|x|_1$ 是凸函数，可以求解，并且保留了解的稀疏性。

数值实验结果

参考 sklearn 举例

SVM 分类器: sklearn SVM 举例
压缩感知: sklearn 压缩感知举例

谢谢大家!