支持向量机与核算法

刘新旺 教授

https://xinwangliu.github.io/

国防科技大学 计算机学院

计算科学系人工智能与大数据教研室

2020 年 10 月 20 日

Content

1. 线性分类
2. 支持向量机
3. 总结

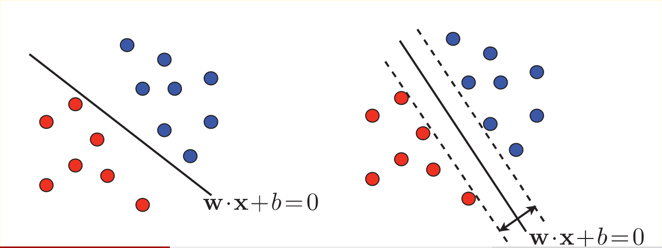
分类：日常生活中的例子

- 信用卡申请Credit card application
- 垃圾邮件过滤
- 网络入侵检测
- …

分类：日常生活中的例子

- 信用卡申请Credit card application
- 垃圾邮件过滤
- 网络入侵检测
- ...

示例



基本符号

- $\mathcal{X} \subseteq \mathbb{R}^d$: 输入空间
- $\mathcal{Y} = \{-1, +1\}$: 输出空间
- $f : \mathcal{X} \mapsto \mathcal{Y}$: 未知的目标函数

- $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$，其中$y_i = f(\mathbf{x}_i)$，$\{\mathbf{x}_i\}_{i=1}^n$从输入空间$\mathcal{X}$中依据某个分布$\mathcal{D}$采集得到：训练集合
- $\mathcal{H} = \{\mathbf{x} \mapsto \mathrm{sgn}(\boldsymbol{\omega}^\top \mathbf{x} + b) : \boldsymbol{\omega} \in \mathbb{R}^d, b \in \mathbb{R}\}$: 假说集合
- $R_{\mathcal{S}}(h) = \frac{1}{n} \sum_{i=1}^n [h(\mathbf{x}_i) \neq f(\mathbf{x}_i)]$ with $h \in \mathcal{H}$: 经验误差
- $R_{\mathcal{D}}(h) = \Pr_{\mathbf{x} \sim \mathcal{D}} [h(\mathbf{x}) \neq f(\mathbf{x})]$ with $h \in \mathcal{H}$: 泛化误差

二分类问题形式化定义

给定训练样本集合 $\mathcal{S} = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\}$，其中$\{\mathbf{x}_i\}_{i=1}^n \subseteq \mathcal{X}^n$ 独立同分布，$y_i = f(\mathbf{x}_i) \in \mathcal{Y} \, (\forall i = 1, \cdots, n)$。二分类问题的目标是基于数据 $\mathcal{S}$，从假说集合 $\mathcal{H}$ 中选择一个假说 h，以使得期望误差

$$\mathrm{E}_{\mathbf{x}\sim\mathcal{D}} \left[h(\mathbf{x}) \neq f(\mathbf{x}) \right] \tag{1}$$

最小。

二分类问题形式化定义

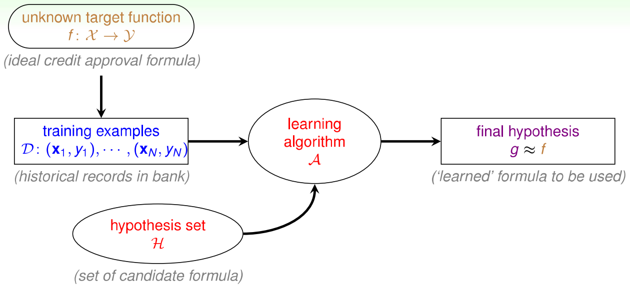
给定训练样本集合 $\mathcal{S} = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\}$，其中$\{\mathbf{x}_i\}_{i=1}^n \subseteq \mathcal{X}^n$ 独立同分布，$y_i = f(\mathbf{x}_i) \in \mathcal{Y}\ (\forall i = 1, \cdots, n)$。二分类问题的目标是基于数据 $\mathcal{S}$，从假说集合 $\mathcal{H}$ 中选择一个假说 h，以使得期望误差

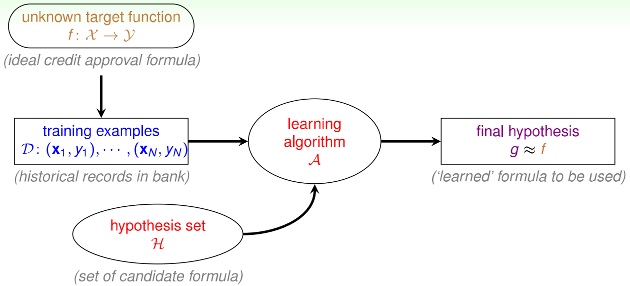
$$\mathrm{E}_{\mathbf{x}\sim\mathcal{D}}\left[h(\mathbf{x}) \neq f(\mathbf{x})\right] \tag{1}$$

最小。

$$\begin{aligned}\mathrm{E}_{\mathbf{x}\sim\mathcal{D}}\left[h(\mathbf{x}) \neq f(\mathbf{x})\right] &= 1 \cdot \Pr_{\mathbf{x}\sim\mathcal{D}}(h(\mathbf{x}) \neq f(\mathbf{x})) + 0 \cdot \Pr_{\mathbf{x}\sim\mathcal{D}}(h(\mathbf{x}) = f(\mathbf{x})) \\ &= \Pr_{\mathbf{x}\sim\mathcal{D}}(h(\mathbf{x}) \neq f(\mathbf{x}))\end{aligned} \tag{2}$$

学习一个分类器

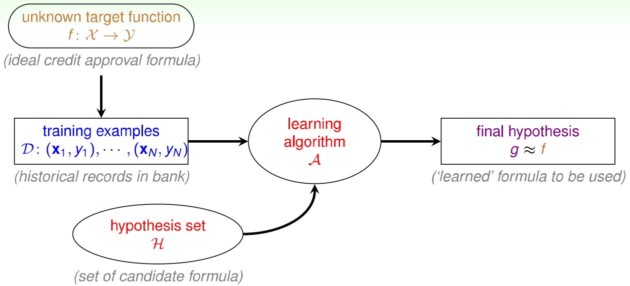

unknown target function
$f\colon \mathcal{X} \to \mathcal{Y}$
(ideal credit approval formula)
training examples
$\mathcal{D}\colon (\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_N, y_N)$
(historical records in bank)
learning
algorithm
$\mathcal{A}$
final hypothesis
$g \approx f$
('learned' formula to be used)
hypothesis set
$\mathcal{H}$
(set of candidate formula)



假说集(线性)

$$\mathcal{H} = \{\mathbf{x} \mapsto \operatorname{sgn}(\boldsymbol{\omega}^\top \mathbf{x} + b) : \boldsymbol{\omega} \in \mathbb{R}^d, b \in \mathbb{R}\}.$$

学习一个分类器



假说集(线性)

$$\mathcal{H} = \{\mathbf{x} \mapsto \operatorname{sgn}(\boldsymbol{\omega}^\top \mathbf{x} + b) : \boldsymbol{\omega} \in \mathbb{R}^d, b \in \mathbb{R}\}.$$

学习算法

$\mathcal{A}$: 支持向量机（Support Vector Machines，SVMs）

假说集

$\mathcal{H} = \{\mathbf{x} \mapsto \text{sgn}[\boldsymbol{\omega}^\top \mathbf{x} + b] : \boldsymbol{\omega} \in \mathbb{R}^d, b \in \mathbb{R}\}.$

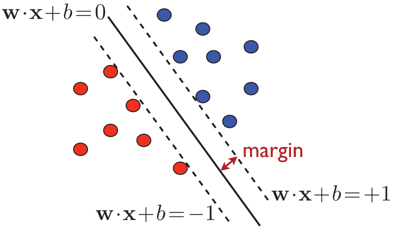
- 假说集 $h \in \mathcal{H} : \mathbf{x} \in \mathcal{X} \mapsto \{-1, +1\}.$
- 决策函数 $h' : \mathbf{x} \in \mathcal{X} \mapsto \boldsymbol{\omega}^\top \mathbf{x} + b \in \mathbf{R}.$
- $h(x) = \text{sgn}\,[h'(x)]$。因为假说$h$由$h'$确定，$h'$是我们研究关注的重点，因此我们通常直接称它为分类器。
- 线性分类器: h'是线性的.

Content

1 线性分类

2 支持向量机
- 线性可分
- Non-Separable Case
- 实现
- 选择正则化参数C

3 总结



(a) 支持向量机

- 考虑超平面$\boldsymbol{\omega}^\top \mathbf{x} + b = 0$
- 给定$a > 0$，我们要求该超平面
  - 对于正类样本(即$y_i = 1$)
    $$\boldsymbol{\omega}^\top \mathbf{x}_i + b \geq a$$
  - 对于负类样本(即$y_i = -1$)
    $$\boldsymbol{\omega}^\top \mathbf{x}_i + b \leq -a$$
  - 即$\boxed{y_i \left(\boldsymbol{\omega}^\top \mathbf{x}_i + b\right) \geq a \, \forall i}$

样本$\mathbf{x}_i$到超平面$\boldsymbol{\omega}^\top \mathbf{x} + b = 0$的距离：$\frac{|\boldsymbol{\omega}^\top \mathbf{x}_i + b|}{\|\boldsymbol{\omega}\|} = \frac{y_i\left(\boldsymbol{\omega}^\top \mathbf{x}_i + b\right)}{\|\boldsymbol{\omega}\|}$

样本集$\mathcal{S}$到超平面$\boldsymbol{\omega}^\top \mathbf{x} + b = 0$的距离$\rho$被定义为：

$$\rho = \min_{(\mathbf{x}_i, y_i) \in \mathcal{S}} \frac{y_i \left(\boldsymbol{\omega}^\top \mathbf{x}_i + b\right)}{\|\boldsymbol{\omega}\|} = \frac{a}{\|\boldsymbol{\omega}\|} \tag{3}$$

优化目标

$$\max_{\boldsymbol{\omega},b} \quad \frac{a}{\|\boldsymbol{\omega}\|} \quad s.t. \quad y_i\left(\boldsymbol{\omega}^\top \mathbf{x}_i + b\right) \geq a, \ \forall i \tag{4}$$

优化目标

$$\max_{\omega,b} \quad \frac{a}{\|\omega\|} \quad s.t. \;\; y_i\left(\omega^\top \mathbf{x}_i + b\right) \geq a, \; \forall i \tag{4}$$

定义$\hat{\omega} = \frac{\omega}{a}$和$\hat{b} = \frac{b}{a}$，EQ.(4)可以等价地写成

优化目标–等价形式

$$\max_{\hat{\omega},\hat{b}} \quad \frac{1}{\|\hat{\omega}\|} \quad s.t. \quad y_i\left(\hat{\omega}^\top \mathbf{x}_i + \hat{b}\right) \geq 1, \; \forall i \tag{5}$$

这种变换($(\omega, b) \Rightarrow (\hat{\omega}, \hat{b})$)会不会影响分类器的预测性能？

优化目标

$$\max_{\boldsymbol{\omega},b} \quad \frac{a}{\|\boldsymbol{\omega}\|} \quad s.t. \;\; y_i\left(\boldsymbol{\omega}^\top \mathbf{x}_i + b\right) \geq a, \; \forall i \tag{4}$$

定义$\hat{\boldsymbol{\omega}} = \frac{\boldsymbol{\omega}}{a}$和$\hat{b} = \frac{b}{a}$，EQ.(4)可以等价地写成

优化目标–等价形式

$$\max_{\hat{\boldsymbol{\omega}},\hat{b}} \quad \frac{1}{\|\hat{\boldsymbol{\omega}}\|} \quad s.t. \quad y_i\left(\hat{\boldsymbol{\omega}}^\top \mathbf{x}_i + \hat{b}\right) \geq 1, \; \forall i \tag{5}$$

这种变换$((\boldsymbol{\omega}, b) \Rightarrow (\hat{\boldsymbol{\omega}}, \hat{b}))$会不会影响分类器的预测性能？

$h(\mathbf{x}) = \mathrm{sgn}(\boldsymbol{\omega}^\top \mathbf{x} + b) = \mathrm{sgn}\left(a\hat{\boldsymbol{\omega}}^\top \mathbf{x} + a\hat{b}\right) = \mathrm{sgn}\left(\hat{\boldsymbol{\omega}}^\top \mathbf{x} + \hat{b}\right) \triangleq \hat{h}(\mathbf{x}) \; (a > 0)$

优化目标—概念上等价

$$\begin{aligned}\min_{\hat{\omega},\hat{b}} \quad & \frac{1}{2}\|\hat{\omega}\|^2 \\ s.t. \quad & y_i\left(\hat{\omega}^\top \mathbf{x}_i + \hat{b}\right) \geq 1, \ \forall i\end{aligned} \tag{6}$$

- 优化目标是严格正定的（Hessian矩阵）
- 所有的约束都是线性的
- EQ.(6)中的优化问题是具有线性约束的二次规划—典型的凸优化问题。
- 最优解存在且唯一

Lagrange函数

$$\mathcal{L}(\hat{\boldsymbol{\omega}}, \hat{b}; \boldsymbol{\alpha}) = \frac{1}{2}\|\hat{\boldsymbol{\omega}}\|^2 - \sum_{i=1}^{n} \alpha_i \left(y_i \left(\hat{\boldsymbol{\omega}}^\top \mathbf{x}_i + \hat{b} \right) - 1 \right) \tag{7}$$

其中$\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_n]^\top$, $\alpha_i \geq 0\,(\forall i)$ 为Lagrange乘子。

KKT条件

$$\begin{cases} \nabla_{\hat{\boldsymbol{\omega}}}\mathcal{L} = \hat{\boldsymbol{\omega}} - \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i = 0 & \Longrightarrow \hat{\boldsymbol{\omega}} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \\ \nabla_{\hat{b}}\mathcal{L} = -\sum_{i=1}^{n} \alpha_i y_i = 0 & \Longrightarrow \sum_{i=1}^{n} \alpha_i y_i = 0 \\ \forall i \;\; \alpha_i \left(y_i \left(\hat{\boldsymbol{\omega}}^\top \mathbf{x}_i + \hat{b} \right) - 1 \right) = 0 & \Longrightarrow \alpha_i \vee y_i \left(\hat{\boldsymbol{\omega}}^\top \mathbf{x}_i + \hat{b} \right) = 1 \end{cases} \tag{8}$$

对偶问题

$$
\begin{aligned}
&\min_{\boldsymbol{\alpha}} \ \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j\left(\mathbf{x}_i^\top\mathbf{x}_j\right) - \sum_{i=1}^{n}\alpha_i \\
&s.t. \ \sum_{i=1}^{n}\alpha_i y_i = 0,\ \alpha_i \geq 0,\ \forall i.
\end{aligned}
\tag{9}
$$

上式是一个标准的带约束二次规划问题。

- 支持向量(Support Vectors): 对应$\alpha_i \geq 0\,(\forall i)$的样本
- 怎样计算$\hat{b}$?

对偶问题

$$\min_{\boldsymbol{\alpha}} \ \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \left(\mathbf{x}_i^\top \mathbf{x}_j\right) - \sum_{i=1}^{n}\alpha_i$$
$$s.t. \ \sum_{i=1}^{n}\alpha_i y_i = 0, \ \alpha_i \geq 0, \ \forall i. \tag{9}$$

上式是一个标准的带约束二次规划问题。

- 支持向量(Support Vectors): 对应$\alpha_i \geq 0\ (\forall i)$的样本
- 怎样计算$\hat{b}$? $\hat{b} = y_i - \sum_{j=1}^{n}\alpha_j y_j \left(\mathbf{x}_j^\top \mathbf{x}_i\right)\ (\alpha_i > 0)$

对偶问题

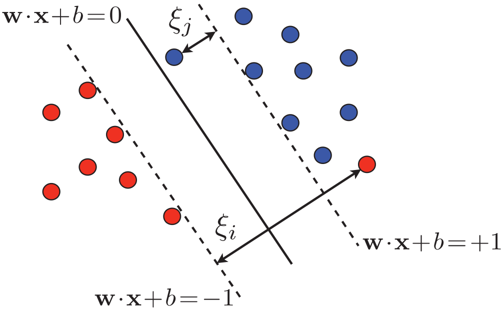
$$\min_{\boldsymbol{\alpha}} \ \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \left(\mathbf{x}_i^\top \mathbf{x}_j\right) - \sum_{i=1}^{n}\alpha_i$$
$$s.t. \ \sum_{i=1}^{n}\alpha_i y_i = 0, \ \alpha_i \geq 0, \ \forall i. \tag{9}$$

上式是一个标准的带约束二次规划问题。

- 支持向量(Support Vectors): 对应$\alpha_i \geq 0 \ (\forall i)$的样本
- 怎样计算$\hat{b}$? $\hat{b} = y_i - \sum_{j=1}^{n}\alpha_j y_j \left(\mathbf{x}_j^\top \mathbf{x}_i\right) \ (\alpha_i > 0)$
- 预测函数：$\hat{h}(\mathbf{x}) = \operatorname{sgn}\left(\hat{\boldsymbol{\omega}}^\top \mathbf{x} + \hat{b}\right) = \operatorname{sgn}\left(\sum_{j=1}^{n}\alpha_j y_j \left(\mathbf{x}_j^\top \mathbf{x}\right) + \hat{b}\right)$
- 样本总是成对出现

- 在绝大多数应用中，训练数据并非线性可分，即对任意超平面 $\boldsymbol{\omega}^\top \mathbf{x} + b = 0$，存在$\mathbf{x}_i \in \mathcal{S}$，使得

$$y_i \left(\boldsymbol{\omega}^\top \mathbf{x}_i + b \right) \ngeq 1 \tag{10}$$



超平面$\boldsymbol{\omega}^\top \mathbf{x} + b = 0$: i)错误地分类$\mathbf{x}_i$; ii)正确地分类$\mathbf{x}_j$，但是间隔小于1

非线性可分—优化目标

考虑如下两个相互矛盾的因素

- 间距最大化
- 训练误差最小化

考虑如下两个相互矛盾的因素

- 间距最大化
- 训练误差最小化

优化目标—松弛变量

$$\min_{\hat{\omega},\hat{b},\boldsymbol{\xi}} \quad \frac{1}{2}\|\hat{\omega}\|^2 + C\sum_{i=1}^{n}\xi_i \quad s.t. \quad y_i\left(\hat{\omega}^\top \mathbf{x}_i + \hat{b}\right) \geq 1 - \xi_i,\ \xi_i \geq 0,\ \forall i \tag{11}$$

- 允许训练过程中有误差
- C为正则化参数—调节模型复杂度与训练误差

$$\begin{aligned}\mathcal{L}(\hat{\boldsymbol{\omega}}, \hat{b}, \boldsymbol{\xi}; \boldsymbol{\alpha}, \boldsymbol{\beta}) &= \frac{1}{2}\|\hat{\boldsymbol{\omega}}\|^2 + C\sum_{i=1}^{n}\xi_i \\ &\quad - \sum_{i=1}^{n}\alpha_i\left(y_i\left(\hat{\boldsymbol{\omega}}^\top \mathbf{x}_i + \hat{b}\right) - 1 + \xi_i\right) - \sum_{i=1}^{n}\beta_i\xi_i\end{aligned} \tag{12}$$

其中$\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_n]^\top$, $\boldsymbol{\beta} = [\beta_1, \cdots, \beta_n]^\top$ 为Lagrange乘子。

KKT条件

$$\begin{cases} \nabla_{\hat{\boldsymbol{\omega}}}\mathcal{L} = \hat{\boldsymbol{\omega}} - \sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i = 0 & \Longrightarrow \hat{\boldsymbol{\omega}} = \sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i \\ \nabla_{\hat{b}}\mathcal{L} = -\sum_{i=1}^{n}\alpha_i y_i = 0 & \Longrightarrow \sum_{i=1}^{n}\alpha_i y_i = 0 \\ \nabla_{\xi_i}\mathcal{L} = C - \alpha_i - \beta_i & \Longrightarrow \alpha_i + \beta_i = C \\ \forall i \;\; \alpha_i\left(y_i\left(\hat{\boldsymbol{\omega}}^\top \mathbf{x}_i + \hat{b}\right) - 1 + \xi_i\right) = 0 & \Longrightarrow \alpha_i \vee y_i\left(\hat{\boldsymbol{\omega}}^\top \mathbf{x}_i + \hat{b}\right) = 1 - \xi_i \\ \forall i \;\; \beta_i\xi_i = 0 & \Longrightarrow \beta_i = 0 \vee \xi_i = 0 \end{cases} \tag{13}$$

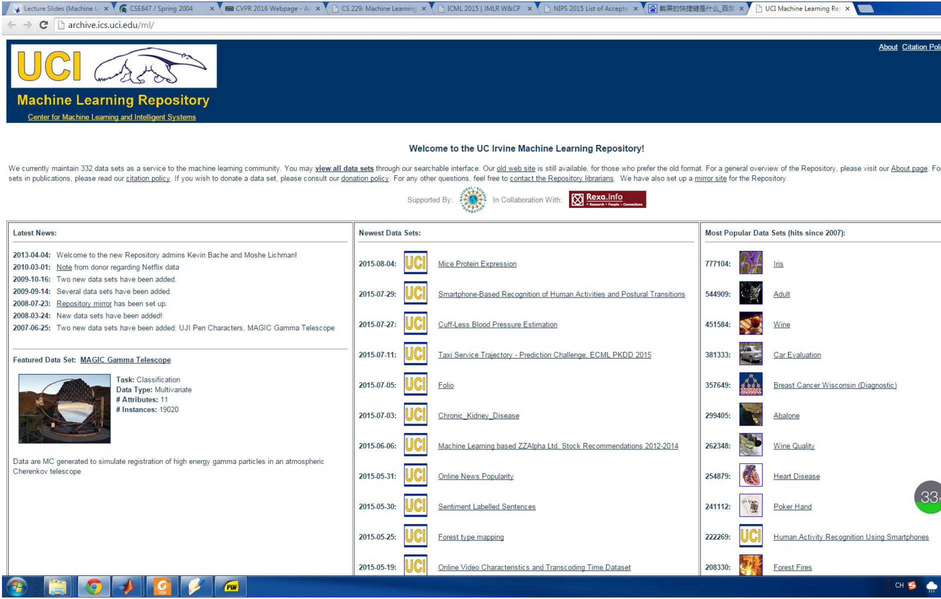
对偶问题

$$
\begin{aligned}
&\min_{\boldsymbol{\alpha}} \ \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \left(\mathbf{x}_i^\top \mathbf{x}_j\right) - \sum_{i=1}^{n}\alpha_i \\
&s.t. \ \sum_{i=1}^{n}\alpha_i y_i = 0, \ C \geq \alpha_i \geq 0, \ \forall i.
\end{aligned}
\tag{14}
$$

```
function [w,d,pos,obj]=mySVMclass(y,c,X)

if min(y) ~= -1
    error(' y must coded: 1 for class one and -1 for class two')
end
%-------------------------------------------------------------------
%      monqp(H,b,c) solves the quadratic programming problem:
%
%    min 0.5*x'Hx - d'x   subject to:  A'x = b  and  0 <= x <= c
%     x
%-------------------------------------------------------------------
K = X*X';
KY = sparse(y*y');
H = K.*KY;
e = ones(size(y));
A = y;
b = 0;
[alpha,lambda0, pos] = mymonqp(H,e,A,b,c);
alphaall=zeros(size(e));
alphaall(pos)=alpha;
obj=-0.5*alphaall'*H*alphaall +e'*alphaall;
ysup = y(pos);
w = (alpha.*ysup);
d = lambda0;
```

UCI数据集



archive.ics.uci.edu/ml/
About Citation Poli
UCI
Machine Learning Repository
Center for Machine Learning and Intelligent Systems
Welcome to the UC Irvine Machine Learning Repository!
We currently maintain 332 data sets as a service to the machine learning community. You may view all data sets through our searchable interface. Our old web site is still available, for those who prefer the old format. For a general overview of the Repository, please visit our About page. Fo
sets in publications, please read our citation policy. If you wish to donate a data set, please consult our donation policy. For any other questions, feel free to contact the Repository librarians. We have also set up a mirror site for the Repository.
Supported By:
In Collaboration With:
Rexa.info
Latest News:
2013-04-04: Welcome to the new Repository admins Kevin Bache and Moshe Lichman!
2010-03-01: Note from donor regarding Netflix data
2009-10-16: Two new data sets have been added.
2009-09-14: Several data sets have been added.
2008-07-23: Repository mirror has been set up.
2008-03-24: New data sets have been added!
2007-06-25: Two new data sets have been added: UJI Pen Characters, MAGIC Gamma Telescope
Featured Data Set: MAGIC Gamma Telescope
Task: Classification
Data Type: Multivariate
Attributes: 11
Instances: 19020
Data are MC generated to simulate registration of high energy gamma particles in an atmospheric Cherenkov telescope
Newest Data Sets:
2015-08-04: Mice Protein Expression
2015-07-29: Smartphone-Based Recognition of Human Activities and Postural Transitions
2015-07-27: Cuff-Less Blood Pressure Estimation
2015-07-11: Taxi Service Trajectory - Prediction Challenge, ECML PKDD 2015
2015-07-05: Folio
2015-07-03: Chronic_Kidney_Disease
2015-06-06: Machine Learning based ZZAlpha Ltd. Stock Recommendations 2012-2014
2015-05-31: Online News Popularity
2015-05-30: Sentiment Labelled Sentences
2015-05-25: Forest type mapping
2015-05-19: Online Video Characteristics and Transcoding Time Dataset
Most Popular Data Sets (hits since 2007):
777104: Iris
544909: Adult
451584: Wine
381333: Car Evaluation
357649: Breast Cancer Wisconsin (Diagnostic)
299405: Abalone
262348: Wine Quality
254879: Heart Disease
241112: Poker Hand
222269: Human Activity Recognition Using Smartphones
208330: Forest Fires

留一法（Leave-One-Out，LOO）

- Training set $\mathcal{S} = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\}$
- Approach: Repeatedly leave one example out for testing.

Train	Test
$\{(\mathbf{x}_2, y_2), (\mathbf{x}_3, y_3), \cdots, (\mathbf{x}_n, y_n)\}$	$(\mathbf{x}_1, y_1)$
$\{(\mathbf{x}_1, y_1), (\mathbf{x}_3, y_3), \cdots, (\mathbf{x}_n, y_n)\}$	$(\mathbf{x}_2, y_2)$
$\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_n, y_n)\}$	$(\mathbf{x}_3, y_3)$
$\ldots$	$\ldots$
$\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_{n-1}, y_{n-1})\}$	$(\mathbf{x}_n, y_n)$

- Performance Evaluation: $\mathrm{Err}_{loo}(\mathcal{A}, C) = \frac{1}{n}\sum_{i=1}^{n}\left[h_{\mathcal{S}-\{\mathbf{x}_i\}}(\mathbf{x}_i) \neq y_i\right]$

Content

1. 线性分类

2. 支持向量机

3. 总结

总结

- 支持向量机推导
- 优化算法
- Matlab实现
- 参数选择

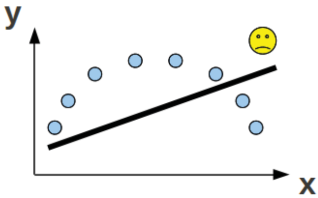
总结

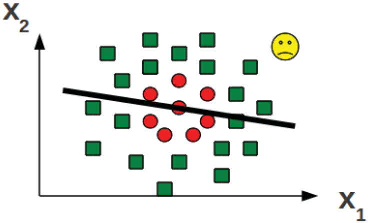
- 支持向量机推导
- 优化算法
- Matlab实现
- 参数选择

Homework 2：实现线性支持向量机算法

支持向量机(线性)优缺点

1. 优点：
   - 算法简单、直观、可理解性强
2. 缺点：表达能力有限，无法捕获数据间的非线性模式。
   - 输入—输出不再是线性关系;
   - 类与类之间不能通过线性边界来划分。





3. Next: Kernel Methods