模式识别引论

An Introduction to Pattern Recognition

主讲: 李春光

www.pris.net.cn/teacher/lichunguang

模式识别与智能系统实验室

网络搜索教研中心 信息与通信工程学院 北京邮电大学

SVM 内容提要

- 引子: 2个类别的分类问题
- 最大间隔(Maximum Margin)分类器
- 支持向量机(SVM)
  - **Support Vector Machine (SVM)**

2个类别的分类问题

- 考虑一个2类分类问题

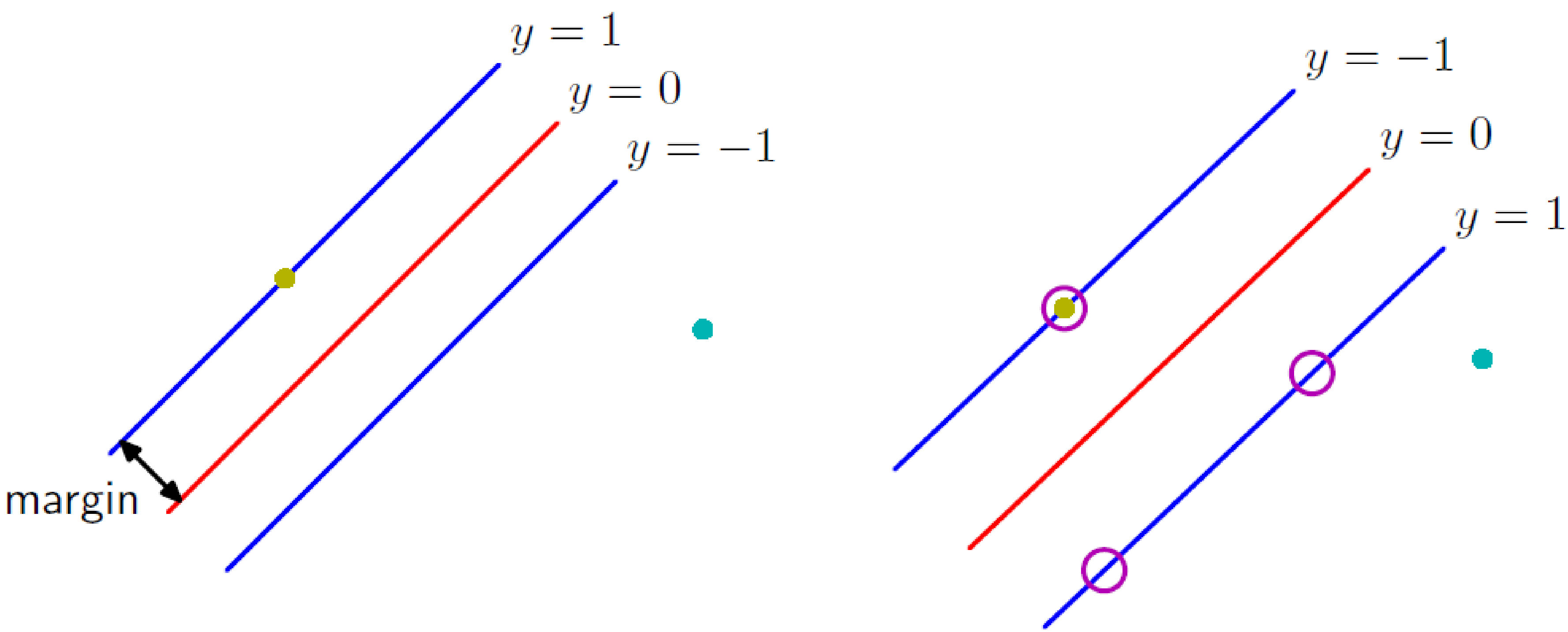
$$y(x) = w^T \phi(x) + b$$

  - 训练数据集
    - N个向量$x_1, \ldots, x_N$
    - N目标输出$t_1, \ldots, t_N$， 二值数据 $t_n \in \{-1, 1\}$
  - 测试数据x
    - 根据 y(x)的符号分类x
  - 假设类别线性可分(Linearly Separable)

    $t_n y(x_n) > 0$

最大间隔分类器

-



Maximum Margin

点x到超平面的距离

- 数据点x到超平面$y(\mathrm{x}) = 0$的距离为

$$|y(\mathrm{x})| / ||\mathrm{w}||$$

- 训练数据点x_n到决策超平面$y(\mathrm{x}) = 0$的距离可以表示为

$$\frac{t_n y(\mathbf{x}_n)}{\|\mathbf{w}\|} = \frac{t_n(\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n) + b)}{\|\mathbf{w}\|}$$

最大间隔

- 最大间隔(Maximum margin)的定义

$$\arg\max_{\mathbf{w},b}\left\{\frac{1}{\|\mathbf{w}\|}\min_{n}\left[t_n\left(\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)+b\right)\right]\right\}$$

  - 通过放缩(**re-scaling**)权值**w**和**b**的大小,我们可以使得距离超平面最近的点的距离为**1**，即

$$t_n\left(\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)+b\right)=1$$

  - 等价于约束 $t_n\left(\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)+b\right)\geqslant 1 \quad n=1,\ldots,N$
  - 优化问题变为 **maximize** $\|\mathbf{w}\|^{-1}$

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

二次规划问题(**Quadratic Programming)**

最大间隔问题的求解

- 拉格朗日乘子法(Lagrange multipliers method)

$$L(\mathbf{w}, b, \mathbf{a}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{n=1}^{N} a_n \{t_n(\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n) + b) - 1\}$$

  - 其中 $\mathbf{a} = (a_1, \ldots, a_N)^{\mathrm{T}}$

  – 计算**L(w, b, a)**相对于 **w** 和**b**的偏导数，并令之为**0**，得出

$$\mathbf{w} = \sum_{n=1}^{N} a_n t_n \phi(\mathbf{x}_n) \qquad 0 = \sum_{n=1}^{N} a_n t_n$$

最大间隔问题的对偶表示

- 在拉格朗日函数中消去w和b，则得到对偶表示

$$\widetilde{L}(\mathbf{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m k(\mathbf{x}_n, \mathbf{x}_m)$$

- 其中 $a_n \geqslant 0 \quad n = 1, \ldots, N$

$$\sum_{n=1}^{N} a_n t_n = 0$$

$$k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^{\mathrm{T}}\phi(\mathbf{x}')$$

仍然是一个二次规划问题 (Quadratic Programming)

最大间隔问题两种表示的比较

- 原始问题(Primal problem)
  - **M**个优化变量
  - 处理新数据点时，使用

$$y(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) + b \qquad \mathbf{w} = \sum_{n=1}^{N} a_n t_n \phi(\mathbf{x}_n)$$

- 对偶问题(Dual problem)
  - **N**个优化变量
  - 对偶表示允许使用**kernels**
  - 处理新数据点时，使用 $y(\mathbf{x}) = \sum_{n=1}^{N} a_n t_n k(\mathbf{x}, \mathbf{x}_n) + b$

  - 当特征空间的维数大于训练样本数目时，即**M > N**时，求解对偶问题更有效**(efficient)**
    - 包含特征空间的维数为无穷的情况

KKT条件

- *Karush-Kuhn-Tucker* (KKT) conditions

$$
\begin{aligned}
a_n &\geqslant 0 \\
t_n y(\mathbf{x}_n) - 1 &\geqslant 0 \\
a_n \{t_n y(\mathbf{x}_n) - 1\} &= 0
\end{aligned}
$$

$\Rightarrow$ $a_n = 0$ or $t_n y(x_n) = 1$

互补松弛条件 (Complementary condition)

- 强对偶理论
  - 原问题的最小值等于对偶问题的最大值

确定偏置参数b

- 根据 $t_n y(x_n) = 1$

$$t_n \left(\sum_{m \in \mathcal{S}} a_m t_m k(\mathbf{x}_n, \mathbf{x}_m) + b \right) = 1$$

  – 得到:

$$b = \frac{1}{N_{\mathcal{S}}} \sum_{n \in \mathcal{S}} \left(t_n - \sum_{m \in \mathcal{S}} a_m t_m k(\mathbf{x}_n, \mathbf{x}_m) \right)$$

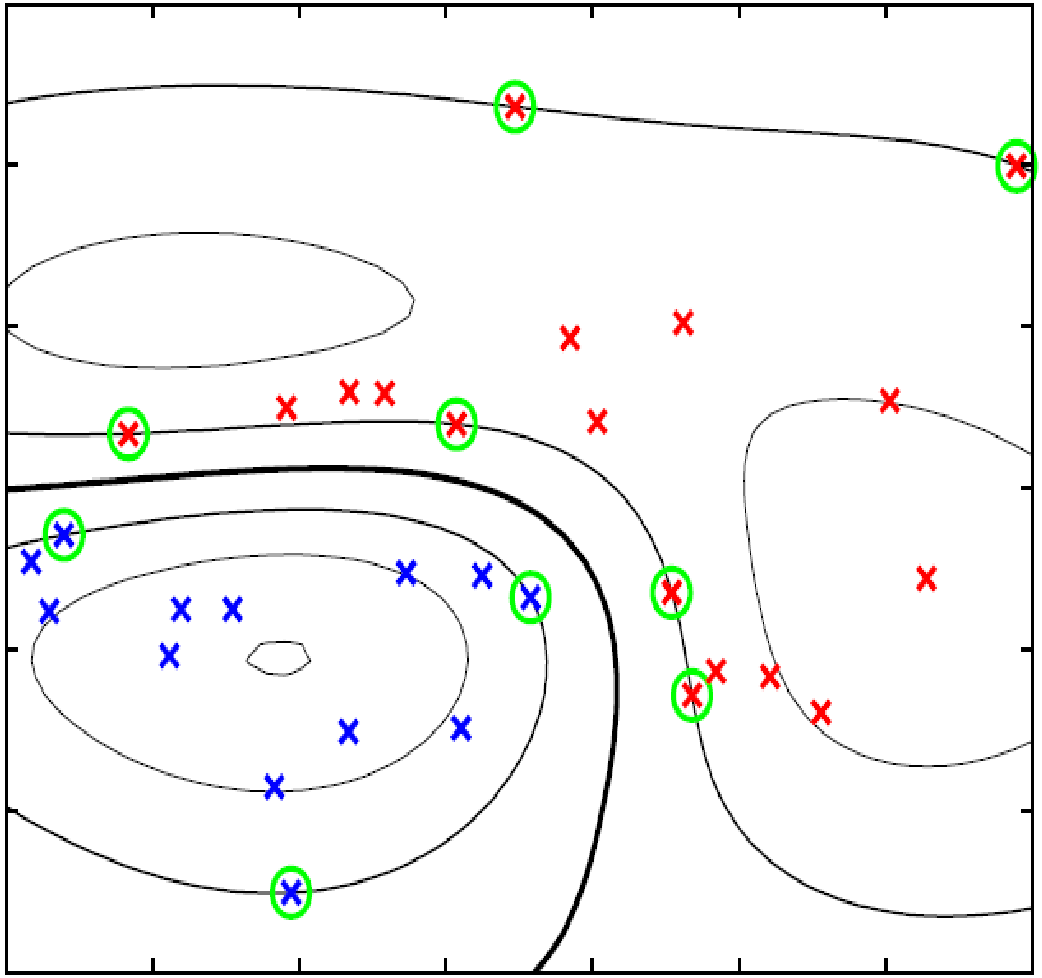
最大间隔问题的另一种表达形式

- 引入误差函数，最大间隔问题的原问题可以改写为

$$\sum_{n=1}^{N} E_{\infty}(y(\mathbf{x}_n)t_n - 1) + \lambda\|\mathbf{w}\|^2$$

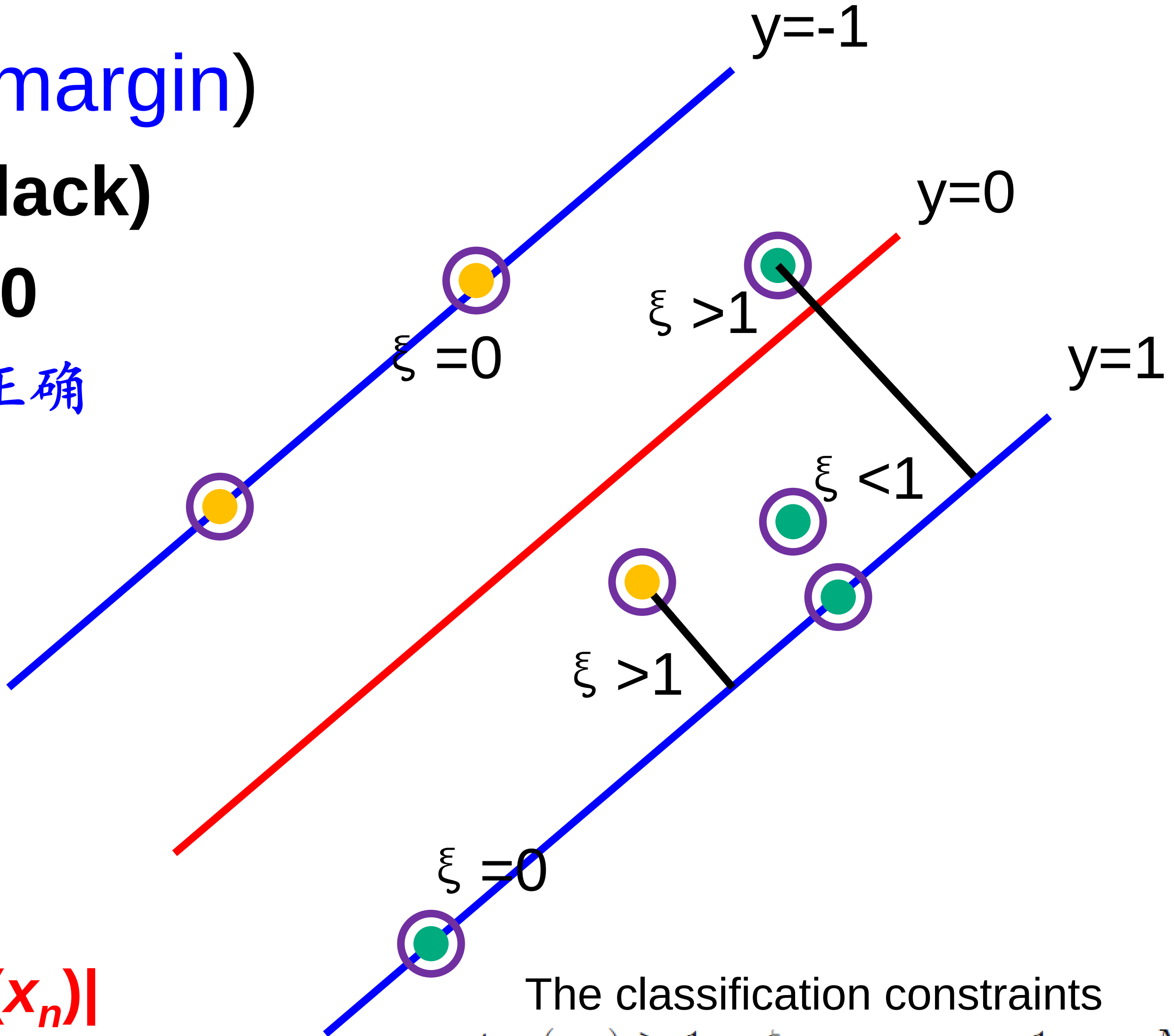
  - 其中 $E_{\infty}(z) = \begin{cases} 0 & z \geq 0 \\ \infty & other \end{cases}$

示例：最大间隔问题

- 

如果类别存在重叠

- 软间隔(soft margin)
  - 引入松弛(slack)变量 $\xi_n \geqslant 0$
    - 对于位于正确的边界之上或之内的数据点有 $\xi_n = 0$
    - 对于其它数据点有 $\xi_n = |t_n - y(x_n)|$



The classification constraints

$$t_n y(\mathbf{x}_n) \geqslant 1 - \xi_n, \qquad n = 1, \ldots, N$$

优化问题转变为…

- 原优化问题变成最小化如下目标函数

$$C\sum_{n=1}^{N}\xi_n + \frac{1}{2}\|\mathbf{w}\|^2$$

  – 其中*C* > *0* 是正则化参数(regularization coefficient)，控制着训练误差与模型复杂度之间的折中

Lagrangian函数变为…

- Lagrangian

$$L(\mathbf{w}, b, \mathbf{a}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{n=1}^{N}\xi_n - \sum_{n=1}^{N} a_n \{t_n y(\mathbf{x}_n) - 1 + \xi_n\} - \sum_{n=1}^{N}\mu_n\xi_n$$

  – **KKT 条件**

$$\begin{aligned} a_n &\geqslant 0 \\ t_n y(\mathbf{x}_n) - 1 + \xi_n &\geqslant 0 \\ a_n \left(t_n y(\mathbf{x}_n) - 1 + \xi_n\right) &= 0 \\ \mu_n &\geqslant 0 \\ \xi_n &\geqslant 0 \\ \mu_n \xi_n &= 0 \end{aligned}$$

where $n = 1, \ldots, N$

寻找对偶问题

- 对w, b, 和 $\{\xi_n\}$求导，令导数为0，则得

$$\frac{\partial L}{\partial \mathbf{w}} = 0 \quad \Rightarrow \quad \mathbf{w} = \sum_{n=1}^{N} a_n t_n \phi(\mathbf{x}_n)$$

$$\frac{\partial L}{\partial b} = 0 \quad \Rightarrow \quad \sum_{n=1}^{N} a_n t_n = 0$$

$$\frac{\partial L}{\partial \xi_n} = 0 \quad \Rightarrow \quad a_n = C - \mu_n.$$

- 带入消去上述变量，则得到Lagrange对偶问题

$$\widetilde{L}(\mathbf{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} a_n a_m t_n t_m k(\mathbf{x}_n, \mathbf{x}_m)$$

  - 其中 $0 \leqslant a_n \leqslant C$ $\qquad \sum_{n=1}^{N} a_n t_n = 0$

box constraints

确定参数b和预测新数据点

- 确定参数b:
  - 对于支持向量(support vectors)，其中**0 < a_n< C** 对应于 ξ $_n$= 0，则有 $t_n y(x_n) = 1$，即满足

$$t_n \left(\sum_{m \in \mathcal{S}} a_m t_m k(\mathbf{x}_n, \mathbf{x}_m) + b \right) = 1$$

  - 从而有

$$b = \frac{1}{N_{\mathcal{M}}} \sum_{n \in \mathcal{M}} \left(t_n - \sum_{m \in \mathcal{S}} a_m t_m k(\mathbf{x}_n, \mathbf{x}_m) \right)$$

- 预测新数据点 $y(\mathbf{x}) = \sum_{n=1}^{N} a_n t_n k(\mathbf{x}, \mathbf{x}_n) + b$

对于解向量的解释

- $a_n = 0$，对应于非支持向量(non-support vectors)
- $a_n > 0$, 对应于支持向量(support vectors), 其中
  - $a_n < C$ $\xi_n = 0$ 位于margin上
  - $a_n = C$ $\xi_n \leqslant 1$ 位于margin之内侧

    $\xi_n > 1$ 误分类(misclassified)

求解SVM中的QP问题

- chunking (Vapnik, 1982)
- protected conjugate gradients (Burges, 1998)
- Decomposition methods (Osuna et al., 1996)
- sequential minimal optimization, or SMO (Platt, 1999)

内积与维数灾难

•

$$
\begin{aligned}
k(\mathbf{x},\mathbf{z}) &= \left(1+\mathbf{x}^{\mathrm{T}}\mathbf{z}\right)^2 = (1+x_1z_1+x_2z_2)^2 \\
&= 1+2x_1z_1+2x_2z_2+x_1^2z_1^2+2x_1z_1x_2z_2+x_2^2z_2^2 \\
&= (1,\sqrt{2}x_1,\sqrt{2}x_2,x_1^2,\sqrt{2}x_1x_2,x_2^2)(1,\sqrt{2}z_1,\sqrt{2}z_2,z_1^2,\sqrt{2}z_1z_2,z_2^2)^{\mathrm{T}} \\
&= \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{z})
\end{aligned}
$$

与Logistic回归的关系

- 最大间隔分类器

$$\min \quad C\sum_{n=1}^{N}\xi_n + \frac{1}{2}\|\mathbf{w}\|^2$$

  – 等价于 $\min \sum_{n=1}^{N} E_{\mathrm{SV}}(y_n t_n) + \lambda\|\mathbf{w}\|^2$

    - 其中

$$E_{\mathrm{SV}}(y_n t_n) = [1 - y_n t_n]_+ \qquad \lambda = (2C)^{-1}$$

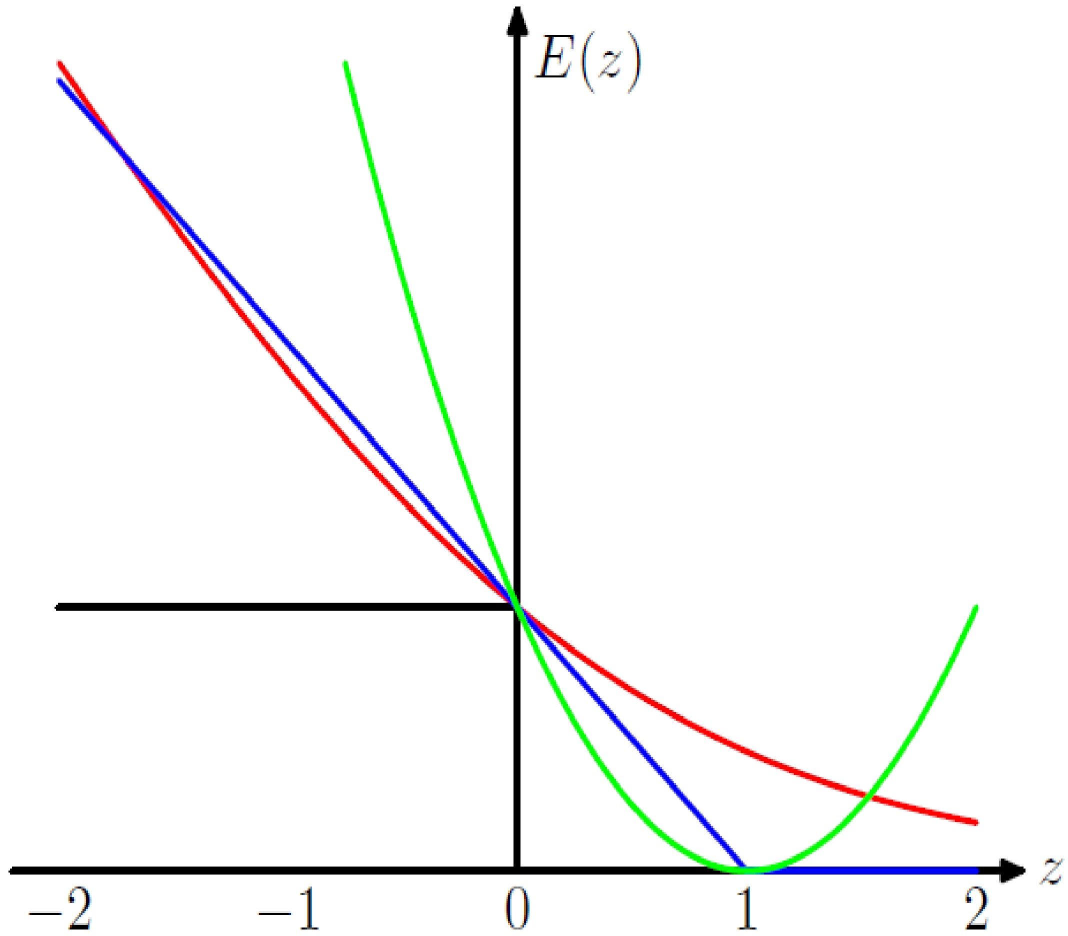
- Logistic回归模型

$$\sum_{n=1}^{N} E_{\mathrm{LR}}(y_n t_n) + \lambda\|\mathbf{w}\|^2$$

    - 其中

$$E_{\mathrm{LR}}(yt) = \ln\left(1 + \exp(-yt)\right)$$

示例: 与Logistic回归的关系

- 

多类情况

- Multiclass SVMs
  - ***one-versus-the-rest* approach**
    - **K-1**
  - ***one-versus-one***
    - **K(K-1)/2**
  - ***single-class***

统计学习理论

- **Statistical Learning Theory**

- 也叫计算学习理论(Computational learning theory)
  - **Anthony and Biggs, 1992; Kearns and Vazirani, 1994; Vapnik, 1995; Vapnik, 1998**
  - **Origin**
    - ***probably approximately correct,* or PAC**
    - **The goal of the PAC framework is to understand how large a data set needs to be in order to give good generalization**

Q / A

- Any Questions...