机器学习与数据科学

Machine Learning and Data Science

主讲: 李春光

www.pris.net.cn/teacher/lichunguang

模式识别与智能系统实验室

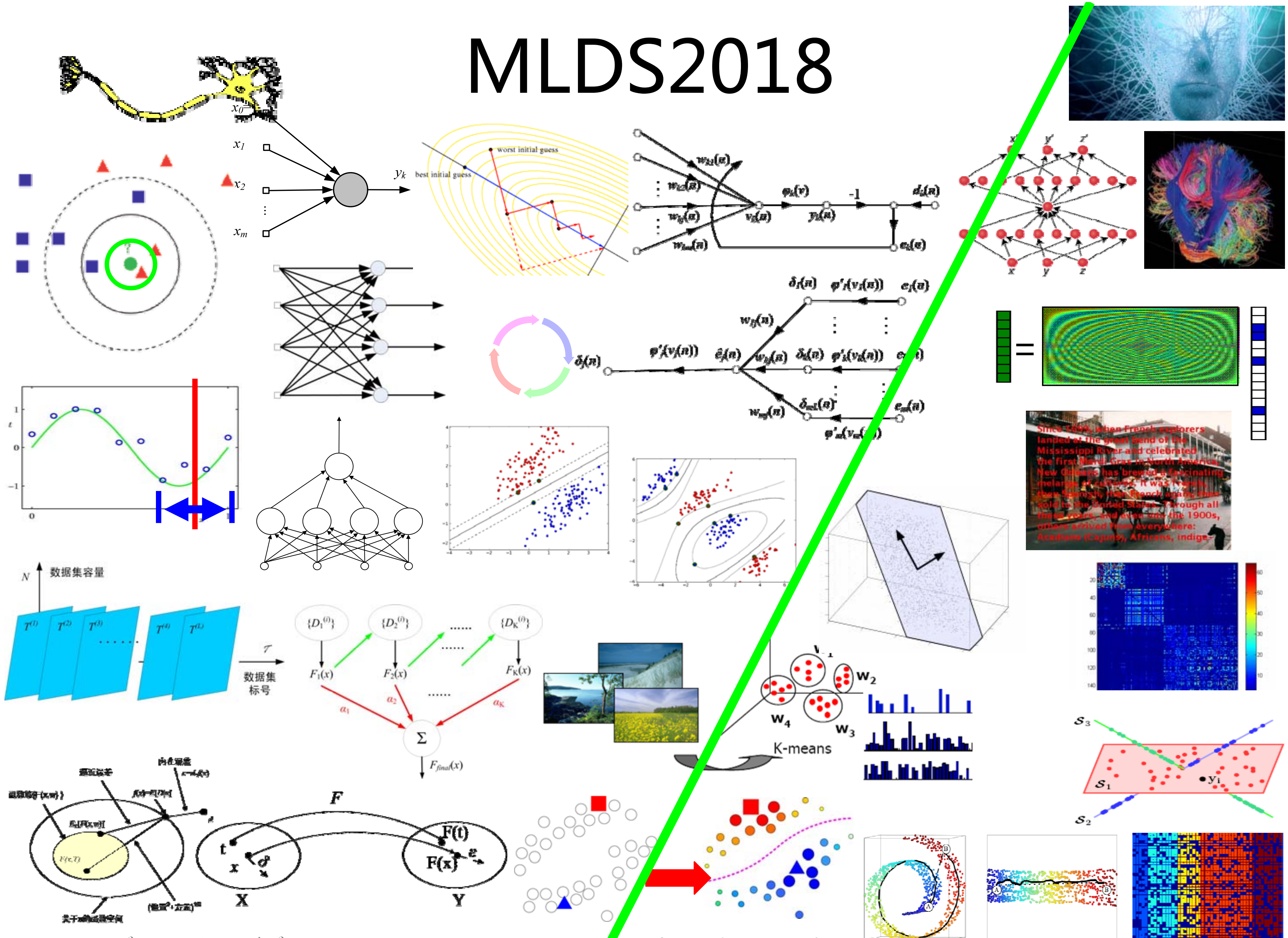
信息与通信工程学院 网络搜索教研中心

北京邮电大学

专题 六：支持向量机与统计学习理论

- **内容提要**
  - 引言
  - 从感知器到支持向量机
  - 统计学习理论
    - 经验风险最小化
    - 结构风险最小化
  - 从结构风险最小化到支持向量机(SVM)

MLDS2018



- **内容提要**
  – 从感知器到支持向量机
  – 统计学习理论
    - 经验风险最小化
    - 结构风险最小化
  – 从结构风险最小化到支持向量机(**SVM**)

类别“线性可分”假设

- 假设：感知器的输入数据来自于两个线性可分的类别 C_1 和 C_2
  - 两个类别“线性可分”，即存在一个权值向量 $\mathbf{w}$，满足：
    - **对于来自类别 C_1 的输入向量 $\mathbf{x}_i$：** $\mathbf{w}^T\mathbf{x}_i + b \ge 0$
    - **对于来自类别 C_2 的输入向量 $\mathbf{x}_i$：** $\mathbf{w}^T\mathbf{x}_i + b < 0$
  - 定义线性决策函数 $g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$
    - **对于样本x，如果** $g(\mathbf{x}) \ge 0$ **，则把x分类到类别1**

      如果 $g(\mathbf{x}) < 0$ **，则把x分类到类别2**

对于样本**x**：如果**g(x)>0**，则**g(x)**越大，分类为$\mathbf{C_1}$越可信；
如果**g(x)<0**，则**g(x)**越小，分类为$\mathbf{C_2}$越可信.

➢ g(x)几何意义是什么呢？

判别函数的几何意义

- 设 $\mathbf{w}$和 b 表示权值向量和偏置，那么决策超平面为

$$\Pi : \mathbf{w}^T \mathbf{x} + b = 0$$

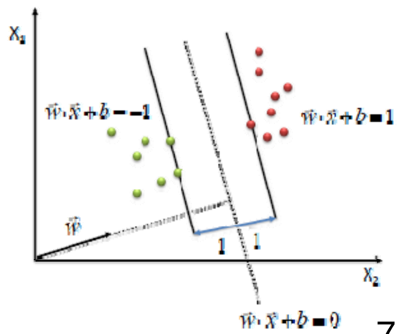
  - 令判别函数 $g(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b$ ，则 $g(\mathbf{x})$给出x到决策超平面距离的代数度量

- 从几何上看：
  - 点x到超平面的代数距离: $\dfrac{1}{\|\mathbf{w}\|_2} g(\mathbf{x})$
  - 原点到超平面的代数距离: $\dfrac{1}{\|\mathbf{w}\|_2} g(0) = \dfrac{b}{\|\mathbf{w}\|_2}$

几何分类间隔与最大间隔超平面

- 给定训练样本 $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ ，假设线性可分 $\begin{cases} y_i = +1 : C_1 \\ y_i = -1 : C_2 \end{cases}$
- **决策超平面方程为：** $g(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b = 0$
- 定义几何分类间隔： $\gamma_i = \dfrac{1}{\|\mathbf{w}\|_2} \cdot y_i \cdot \left(\mathbf{w}^T \mathbf{x}_i + b\right)$
  – 若分类正确，则 $\gamma_i > 0$
- 最优超平面，即具有最大分类间隔的超平面

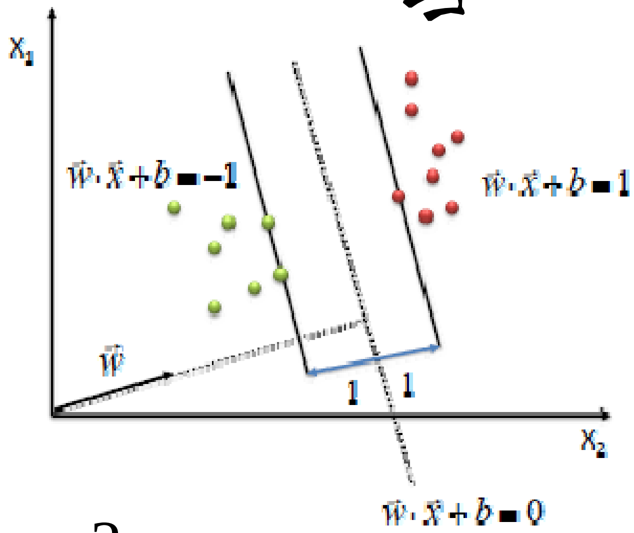
$$\left(\mathbf{w}, b\right) = \arg\max_{\mathbf{w},b} \gamma \quad \text{其中，} \quad \gamma = \min_i \gamma_i$$



➢ 如何简化这个max-min优化问题？

支持向量与最优化问题

- 定义支持向量(support vectors) $\mathbf{x}^{(s)}$: 使得下式取等号的 $\mathbf{x}_i$
  - 对于: $y_i = +1$ $\quad \mathbf{w}_o^T \mathbf{x}_i + b_o \geq 1$
  - 对于 $y_i = -1$ $\quad \mathbf{w}_o^T \mathbf{x}_i + b_o \leq -1$



- 考虑支持向量 $\mathbf{x}^{(s)}$到最优超平面的代数距离：

$$r = \frac{g\left(\mathbf{x}^{(s)}\right)}{\|\mathbf{w}_o\|_2} = \begin{cases} \dfrac{1}{\|\mathbf{w}_o\|_2} & \text{if } \; y^{(s)} = +1 \\ -\dfrac{1}{\|\mathbf{w}_o\|_2} & \text{if } \; y^{(s)} = -1 \end{cases}$$

- 令ρ表示类别分离间隔的最优值，则 $\rho = 2r = \dfrac{2}{\|\mathbf{w}_o\|_2}$

- 最大化类别间隔 $\Leftrightarrow$ 最小化权值向量的**欧几里德范数**

$$\max_{\mathbf{w}} \rho = \max_{\mathbf{w}} \frac{2}{\|\mathbf{w}\|_2} \Leftrightarrow \min_{\mathbf{w}} \frac{\|\mathbf{w}\|_2}{2} \Leftrightarrow \min_{\mathbf{w}} \|\mathbf{w}\|_2^2$$

寻找最优超平面的优化问题

- 给定训练样本集 $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ ，找到权值向量和偏置的最优值 $\mathbf{w}_o$ 和 b_o ，使得它们满足如下约束条件：

$$y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 1, \quad \text{for } i = 1, \ldots, N$$

且权值向量最小化如下代价函数

$$\varepsilon(\mathbf{w}, b) = \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

最优化问题:
1. 无约束且可导
2. 有约束但目标函数可导
3. (无约束但不可导)
4. 有约束且目标函数不可导

线性SVM求解下述最优化问题

$$\arg\min_{\mathbf{w}, b} \varepsilon(\mathbf{w}, b) = \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

$$\text{s.t.} \quad \left(\mathbf{w}^T\mathbf{x}_i + b\right) y_i \geq 1, \quad \text{for } i = 1, \ldots, N$$

求解带约束的优化问题

- Lagrange 乘子法：引入辅助变量 $\{\alpha_i \ge 0\}_{i=1}^{N}$

$$L(\mathbf{w},b,\alpha)=\frac{1}{2}\mathbf{w}^T\mathbf{w}-\sum_{i=1}^{N}\alpha_i\left(y_i\left(\mathbf{w}^T\mathbf{x}_i+b\right)-1\right)$$

- 极小值条件：

$$\frac{\partial L(\mathbf{w},b,\alpha)}{\partial \mathbf{w}}=0, \quad \Rightarrow \quad \mathbf{w}=\sum_{i=1}^{N}\alpha_i y_i \mathbf{x}_i,$$

➢ 几何解释：w位于输入数据所张成线性子空间中

$$\frac{\partial L(\mathbf{w},b,\alpha)}{\partial b}=0 \quad \Rightarrow \quad \sum_{i=1}^{N}\alpha_i y_i=0$$

Lagrange对偶问题

- 把极小值条件代入到Lagrange 辅助函数中：

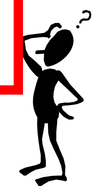
$$\mathbf{w}=\sum_{i=1}^{N}\alpha_i y_i \mathbf{x}_i, \qquad \sum_{i=1}^{N}\alpha_i y_i = 0$$

  – 消掉w和b,

$$D(\boldsymbol{\alpha})=\min_{\mathbf{w},b} L(\mathbf{w},b,\alpha)=\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j,$$

Lagrange对偶问题:

$$\max_{\boldsymbol{\alpha}\ge 0} D(\boldsymbol{\alpha}), \quad \text{where} \quad D(\boldsymbol{\alpha})=\min_{\mathbf{w},b} L(\mathbf{w},b,\boldsymbol{\alpha})$$



求解对偶问题

- 转化为对偶问题(dual problem)：
  - 给定训练样本集 $\left\{\left(\mathbf{x}_i, y_i\right)\right\}_{i=1}^{N}$，**最大化目标函数：**

$$D(\alpha)=\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j$$

满足约束条件：

– **(1)** $\sum_{i=1}^{N}\alpha_i y_i=0$

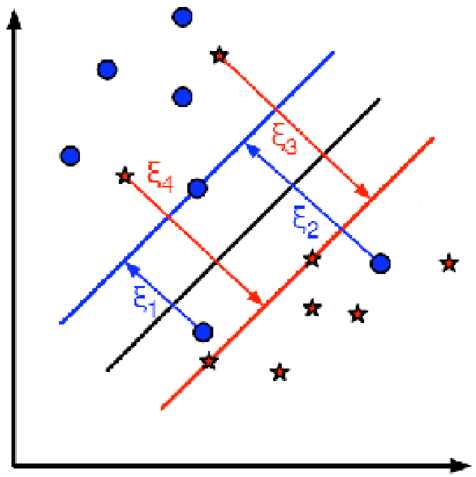
– **(2)** $\alpha_i\geq 0,\ \text{for } i=1,\ldots,N$

最优权值向量 $\mathbf{w}_o$ 为： $\mathbf{w}_o=\sum_{i=1}^{N}\alpha_i^* y_i\mathbf{x}_i$

最优偏置 b_o 为： $b_o=1-\mathbf{w}_o^T\mathbf{x}^{(s)}$ 对于 $y^{(s)}=1$

考虑线性不可分问题

- 定义支持向量(support vectors)$\mathbf{x}^{(s)}$: 使得下式取等号的 $\mathbf{x}_i$
  - 对于 $y_i = +1$ $\quad \mathbf{w}_o^T \mathbf{x}_i + b_o \geq 1$
  - 对于 $y_i = -1$ $\quad \mathbf{w}_o^T \mathbf{x}_i + b_o \leq -1$



- 引入松弛变量**(slack variable)** $\xi_i \geq 0$
  - 对于 $y_i = +1$ $\quad \mathbf{w}_o^T \mathbf{x}_i + b_o \geq 1 - \xi_i$
  - 对于 $y_i = -1$ $\quad \mathbf{w}_o^T \mathbf{x}_i + b_o \leq -1 + \xi_i$
- 目标: 使得类别分离间隔最大，且量度分类错误的松弛变量越小越好

$$\boxed{\begin{cases} \min\limits_{\mathbf{w}} \|\mathbf{w}\|_2^2 \\ \min\limits_{b,\{\xi_i\}_{i=1}^N} \sum_{i=1}^{N} \xi_i \end{cases} \quad \Longleftrightarrow \quad \min_{\mathbf{w},b,\{\xi_i\}_{i=1}^N} \frac{1}{2}\|\mathbf{w}\|_2^2 + C \cdot \sum_{i=1}^{N} \xi_i}$$

求解线性不可分问题的最优超平面

- 给定训练样本集 $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$，找到权值向量和偏置的最优值 $\mathbf{w}_o$ 和 b_o ，使得它们满足如下约束条件：

$$y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots, N$$

且最小化代价函数 $\varepsilon(\mathbf{w}, b, \boldsymbol{\xi}) = \dfrac{1}{2}\mathbf{w}^T\mathbf{w} + C \cdot \sum_{i=1}^{N} \xi_i$

– 即为下述二次规划问题

$$\min_{\mathbf{w}, b, \{\xi_i\}_{i=1}^{N}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}\xi_i$$

$$\text{s.t. } y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots., N$$

求解对偶问题

- 对偶问题(dual problem)：
  - 给定训练样本集$\left\{\left(\mathbf{x}_i, y_i\right)\right\}_{i=1}^{N}$，最大化目标函数：

$$D(\alpha)=\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j$$

满足约束条件：

- (1) $\sum_{i=1}^{N}\alpha_i y_i=0$
- (2) $0\le\alpha_i\le C,\quad i=1,\ldots,N$

最优权值向量$\mathbf{w}_o$为：

$$\mathbf{w}_o=\sum_{i=1}^{N}\alpha_i y_i\mathbf{x}_i$$

比较： 松弛变量以及其Lagrange乘子未出现在对偶问题中，区别是α多了上界约束

非线性核 SVM

- 对偶问题(dual problem)：
  - 给定训练样本集 $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$，最大化目标函数：

$$Q(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$$

满足约束条件：

其中核函数K(.,.)需要指定，比如多项式核、高斯核

  - **(1)** $\sum_{i=1}^{N} \alpha_i y_i = 0$
  - **(2)** $0 \le \alpha_i \le C,\ \text{for } i = 1, \ldots, N$

最优权值向量对应的决策超平面方程为：

$$\sum_{i=1}^{N_S} \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) + b = 0$$

SMO

- 高效求解核SVM:

$$\min_{\{\alpha_i\}_{i=1}^N} \frac{1}{2}\sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j K\left(\mathbf{x}_i, \mathbf{x}_j\right) - \sum_{i=1}^N \alpha_i$$

$$\text{s.t. } \sum_{i=1}^N \alpha_i y_i = 0,\ 0 \le \alpha_i \le C.$$

  – **SMO: Sequential Minimal Optimization**

- 基本思想:
  – 每次求解仅涉及两个优化变量的二次规划问题
    - 在每次迭代时，选中两个优化变量α_{i*} 和 α_{j*} 同时保持其它变量固定，求解关于α_{i*}和α_{j*} 的二次规划问题
      – 在LibSVM中被采用

[1] John Platt: Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines, 1998.

Q / A

- Any Question? …

请思考下面几个问题…

- 为什么要最小化训练错误数？
  - 合理吗？其合理性需要证明
- 为什么要最大化分类间隔?
  - 这一几何直观可以找到理论保证吗? 其合理性需要证明
- 在学习过程中，我们的目标是什么？
  - 获得具有良好泛化能力的学习机器
- 研究学习问题的目标是什么?
  - 寻找能够达到最好的推广性能的归纳原则，并构造算法来实现这一原则
- 如何保证推广性能? 是否有其他的归纳原理，能够达到更好的推广性能？

- **内容提要**
  - 从感知器到支持向量机
  - 统计学习理论
    - 经验风险最小化
    - 结构风险最小化
  - 从结构风险最小化到支持向量机(**SVM**)