Machine Learning and Pattern Recognition

东北大学“智能+X”新工科课程系列

机器学习与模式识别

东北大学 信息科学与工程学院
人工智能系 、智能感知与机器人研究所
陈东岳

Solving Soft Margin SVM

软间隔SVM求解

CHAPTER ONE

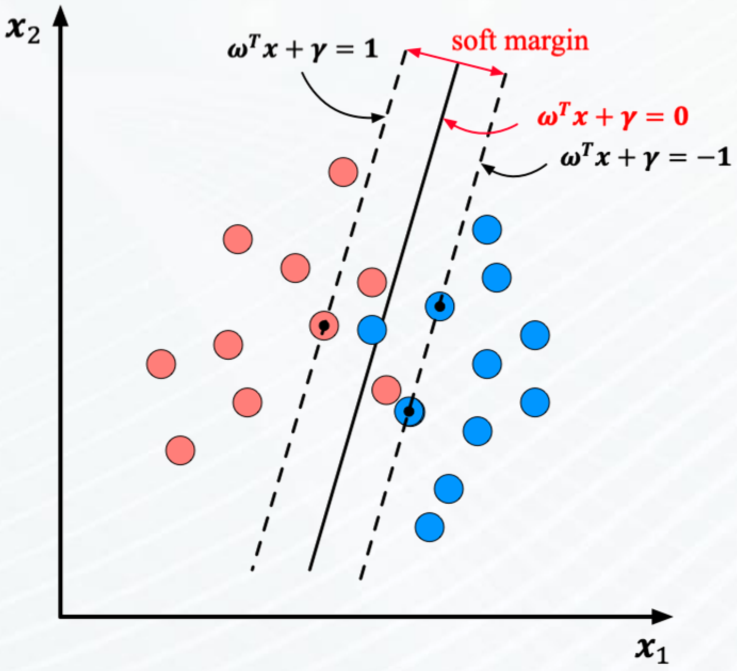
问题模型

Problem model

一、问题模型

▶ 问题的提出——线性不可分

- 大多数分类问题是线性不可分的
- 我们需要一个允许出现分类误差的模型
- 软间隔——样本在一定程度下可以侵入到分类间隔区域内部



一、问题模型

▶ 代理损失(surrogate loss)

- 考虑侵入间隔区域的样本，设计一个新的损失函数。

$$\min_{\boldsymbol{\omega},\gamma} \frac{1}{2}\|\boldsymbol{\omega}\|^2 + C\sum_{i=1}^{m} \ell_{0/1}\left(y_i\left(\boldsymbol{\omega}^T\boldsymbol{x}_i + \gamma\right) - 1\right)$$

- 其中:

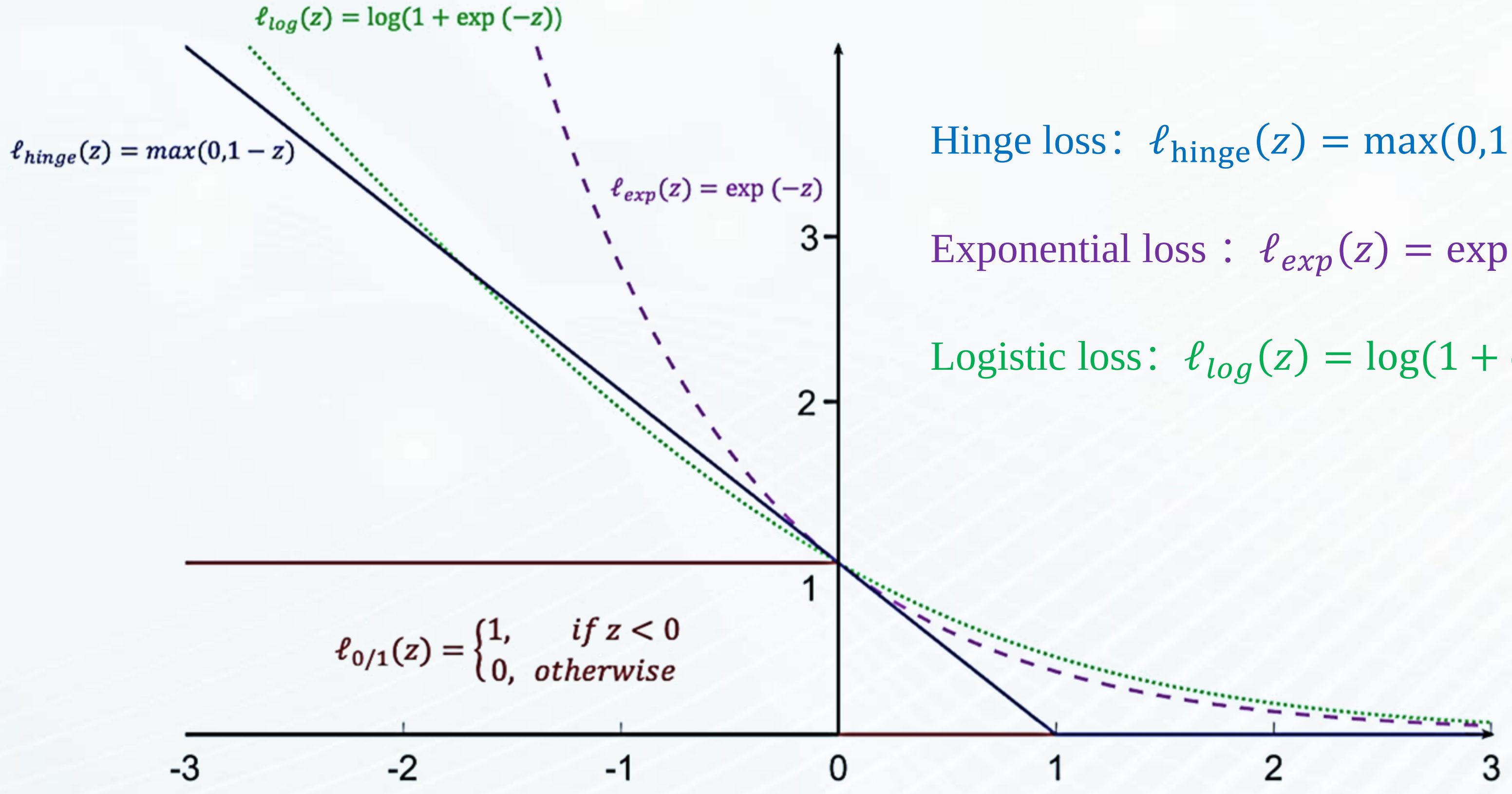
$$\ell_{0/1}(z) = \begin{cases} 1, & \text{if } z < 0 \\ 0, & \text{otherwise} \end{cases}$$

- 然而，$\ell_{0/1}$不是一个足够好的选择，因为$\ell_{0/1}$是不连续的、非凸的函数，因此需要具有更好数学性质的“代理损失函数”

一、问题模型

▶ 代理损失(surrogate loss)

- 常用的代理损失



Hinge loss： $\ell_{\text{hinge}}(z) = \max(0,1-z)$

Exponential loss ： $\ell_{exp}(z) = \exp(-\text{z})$

Logistic loss： $\ell_{log}(z) = \log(1 + \exp(-\text{z}))$

一、问题模型

▶ 松弛变量的引入

Hinge loss： $\ell_{hinge}(z)=\max(0,1-z)$

- 如果采用Hinge loss, 可对每一个训练样本$(\boldsymbol{x}_i, y_i)$引入对应的“松弛变量” ξ_i，此时对应的优化目标为：

$$\min_{\omega,\gamma,\xi} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{m}\xi_i$$

- 相应的约束条件变为：

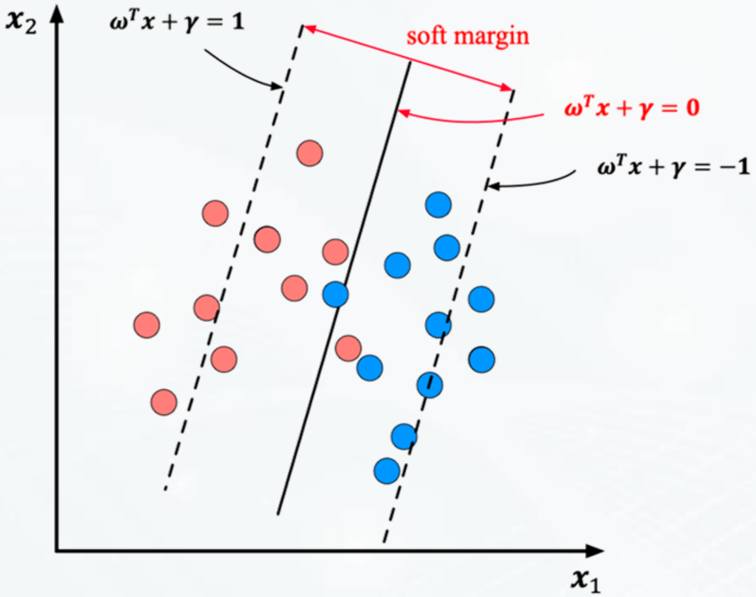
$$y_i\left(\boldsymbol{\omega}^T\boldsymbol{x}_i+\gamma\right)\geq 1-\xi_i$$
$$\xi_i \geq 0, \qquad \forall \boldsymbol{x}_i$$

➢ 引入松弛变量ξ_i，放宽了约束条件，因此可以解决线性不可分问题。

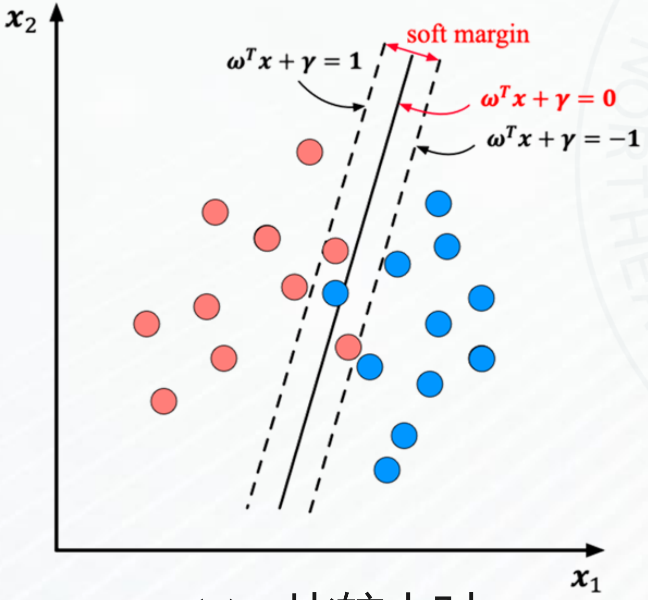
- 思考：目标函数中的C起到了什么作用？

一、问题模型

▶ 惩罚系数C



(1) C比较小时



(2) C比较大时

-C越小，ξ_i越大⇒间隔变大

-C越大，ξ_i越小⇒间隔变小

一、问题模型

▶ 软间隔的原始问题

- 引入松弛变量ξ_i后，软间隔SVM对应的优化问题可以写为：

$$
\begin{aligned}
&\min_{\boldsymbol{\omega},\gamma,\xi} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{m}\xi_i \\
&\text{s. t.} \quad y_i\left(\boldsymbol{\omega}^T\boldsymbol{x}_i + \gamma\right) \geq 1 - \xi_i, \qquad i = 1,2,\dots,m \\
&\qquad\quad \xi_i \geq 0, \qquad\qquad\qquad\qquad i = 1,2,\dots,m
\end{aligned}
$$

- 相应的拉格朗日函数为：

$$
L(\boldsymbol{\omega},\gamma,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\beta}) = \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{m}\xi_i - \sum_{i=1}^{m}\alpha_i\left(y_i\left(\boldsymbol{\omega}^T\boldsymbol{x}_i + \gamma\right) - 1 + \xi_i\right) - \sum_{i=1}^{m}\beta_i\xi_i
$$

其中$\alpha_i \geq 0, \beta_i \geq 0, \forall i=1,2,\ldots,m$

- 则软间隔SVM的原始问题为：

$$
\min_{\boldsymbol{\omega},\gamma,\xi}\left[\max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\alpha_i\geq 0,\beta_i\geq 0} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{m}\xi_i - \sum_{i=1}^{m}\alpha_i\left(y_i\left(\boldsymbol{\omega}^T\boldsymbol{x}_i + \gamma\right) - 1 + \xi_i\right) - \sum_{i=1}^{m}\beta_i\xi_i\right]
$$

CHAPTER TWO

问题转化

Problem transformation

二、问题转化

▶ 软间隔的对偶问题

- 原始问题的拉格朗日对偶问题为：

$$\max_{\alpha,\beta:\alpha_i\geq 0,\beta_i\geq 0}\left[\min_{\omega,\gamma,\xi} L(\boldsymbol{\omega},\gamma,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\beta})\right]$$

其中

$$L(\boldsymbol{\omega},\gamma,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\beta})=\frac{1}{2}\|\boldsymbol{\omega}\|^2+C\sum_{i=1}^{m}\xi_i-\sum_{i=1}^{m}\alpha_i\left(y_i\left(\boldsymbol{\omega}^T\boldsymbol{x}_i+\gamma\right)-1+\xi_i\right)-\sum_{i=1}^{m}\beta_i\xi_i$$

- 求$L(\boldsymbol{\omega},\gamma,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\beta})$对$\boldsymbol{\omega},\gamma,\boldsymbol{\xi}$的极小值，令相应的偏导数为0

$$\frac{\partial L}{\partial \boldsymbol{\omega}}=\mathbf{0}\Longrightarrow \boldsymbol{\omega}=\sum_{i=1}^{m}\alpha_i y_i \boldsymbol{x}_i \qquad (1)$$

$$\frac{\partial L}{\partial \gamma}=0\Longrightarrow 0=\sum_{i=1}^{m}\alpha_i y_i \qquad (2)$$

$$\frac{\partial L}{\partial \boldsymbol{\xi}}=\mathbf{0}\Longrightarrow C-\alpha_i-\beta_i=0,\forall i=1,2,\dots,m \qquad (3)$$

- 将(1)(2)(3)式代入$L(\boldsymbol{\omega},\gamma,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\beta})$中，可得

$$\min L(\boldsymbol{\omega},\gamma,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\beta})=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^T\boldsymbol{x}_j$$

二、问题转化

▶ 软间隔的对偶问题

- 考虑约束条件

$$
\begin{aligned}
s.t. \quad & \sum_{i=1}^{m} \alpha_i y_i = 0 \\
& C - \alpha_i - \beta_i = 0 \\
& \alpha_i \geq 0, i = 1,2,\ldots,m \\
& \beta_i \geq 0, i = 1,2,\ldots,m
\end{aligned}
$$

- 拉格朗日对偶问题可以重写为:

$$
\begin{aligned}
\max_{\boldsymbol{\alpha}} \; & \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j \\
s.t. \quad & \sum_{i=1}^{m} \alpha_i y_i = 0 \\
& 0 \leq \alpha_i \leq C, i = 1,2,\ldots,m
\end{aligned}
$$

二、问题转化

▶ 软间隔的KKT条件

- 软间隔SVM优化问题的KKT条件变为：

$$\begin{cases} \alpha_i \geq 0 & (1) \\ y_i f(\boldsymbol{x}_i) - 1 + \xi_i \geq 0 & (2) \\ \alpha_i(y_i f(\boldsymbol{x}_i) - 1 + \xi_i) = 0 & (3) \\ \beta_i \geq 0 & (4) \\ \xi_i \geq 0 & (5) \\ \beta_i \xi_i = 0 & (6) \end{cases} \quad i = 1,2,\dots,m$$

其中: $f(\boldsymbol{x}) = \boldsymbol{\omega}^T \boldsymbol{x} + \gamma$

二、问题转化

KKT条件的解释

- 对于软间隔KKT条件，可以进行如下解释:

$$\begin{cases} \alpha_i \geq 0, \ \beta_i \geq 0 \\ 1-\xi_i - y_i f(x_i) \leq 0 \\ \alpha_i(1-\xi_i - y_i f(x_i)) = 0 \\ \xi_i \geq 0, \ \beta_i \xi_i = 0 \end{cases} \qquad \begin{cases} \omega = \sum_{i=1}^{m} \alpha_i y_i x_i \\ 0 = \sum_{i=1}^{m} \alpha_i y_i \\ C = \alpha_i + \beta_i \end{cases}$$

- Case1：$\alpha_i = 0$，由 $\begin{cases} C = \alpha_i + \beta_i \\ \beta_i \xi_i = 0 \end{cases} \Rightarrow \begin{cases} \beta_i = C \\ \xi_i = 0 \end{cases} \Rightarrow$ 样本$\boldsymbol{x}_i$位于间隔之外

- Case2：$\alpha_i > 0$，$y_i f(\boldsymbol{x}_i) = 1 - \xi_i$，样本$\boldsymbol{x}_i$是一个支持向量

  - 如果$\alpha_i < C$，那么$\beta_i > 0$， 从而得知$\xi_i = 0$，所以x_i在边界上

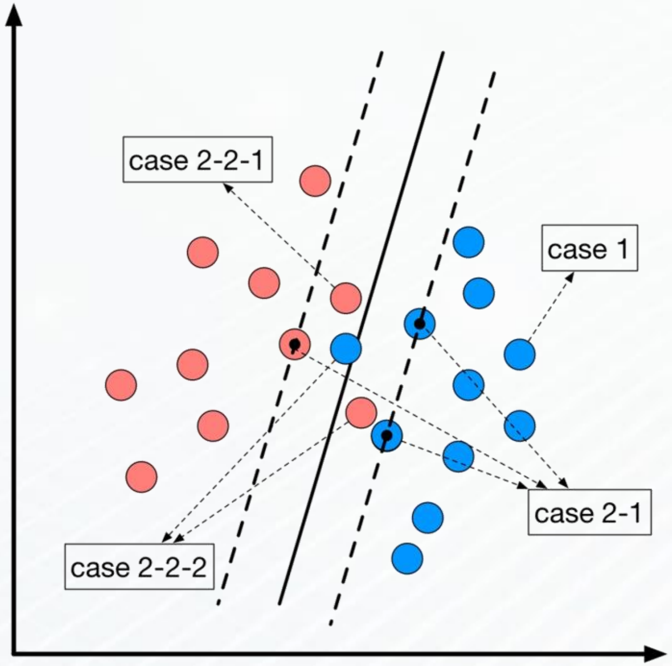
  - 如果$\alpha_i = C$，那么$\beta_i = 0$，

    - 如果$\xi_i \leq 1$，那么x_i在间隔区域内正确的一侧
    - 如果$\xi_i > 1$，那么x_i被错误分类了

二、问题转化

▶KKT条件的解释

- Case1： $\alpha_i = 0$， $\beta_i = C$， $\xi_i = 0$
- Case2： $\alpha_i > 0$， $y_i f(x_i) = 1 - \xi_i$
  - Case2-1： $\alpha_i < C$， $\beta_i > 0$， $\xi_i = 0$
  - Case2-2： $\alpha_i = C$， $\beta_i = 0$，
    - Case2-2-1： $\xi_i \le 1$
    - Case2-2-2： $\xi_i > 1$



CHAPTER THREE

模型解释

Model Explanation

三、模型解释

经验风险与结构风险

- 有别于软间隔SVM的几何解释，还可以对优化问题进行如下改造：

$$\min_{\omega,\gamma} \sum_{i=1}^{m} \ell_h(\boldsymbol{x}_i) + \lambda\|\omega\|^2$$

其中

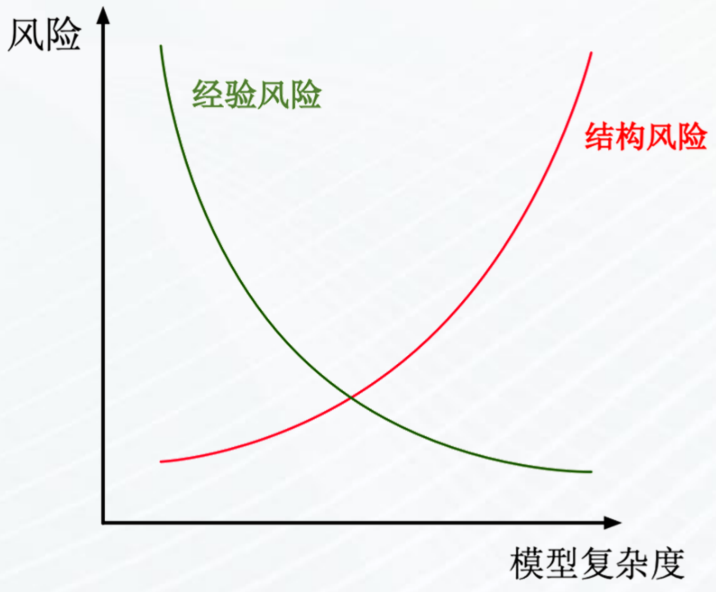
$$\ell_h(\boldsymbol{x}) = \begin{cases} 1 - y(\boldsymbol{\omega}^T\boldsymbol{x} + \gamma), & y(\boldsymbol{\omega}^T\boldsymbol{x} + \gamma) < 1 \\ 0, & y(\boldsymbol{\omega}^T\boldsymbol{x} + \gamma) \geq 1 \end{cases}, \quad \lambda = \frac{1}{2C}$$

- ℓ_h可以用$\ell_{\exp}$或$\ell_{\log}$等代理损失函数来代替，$\sum_{i=1}^{m} \ell_h(\boldsymbol{x}_i)$被称为经验风险（Empirical Risk，ER）
- $\|\omega\|^2$被称为结构风险（Structure Risk，SR）
- λ用来调和ER和SR之间的矛盾。

三、模型解释

▶ 经验风险与结构风险

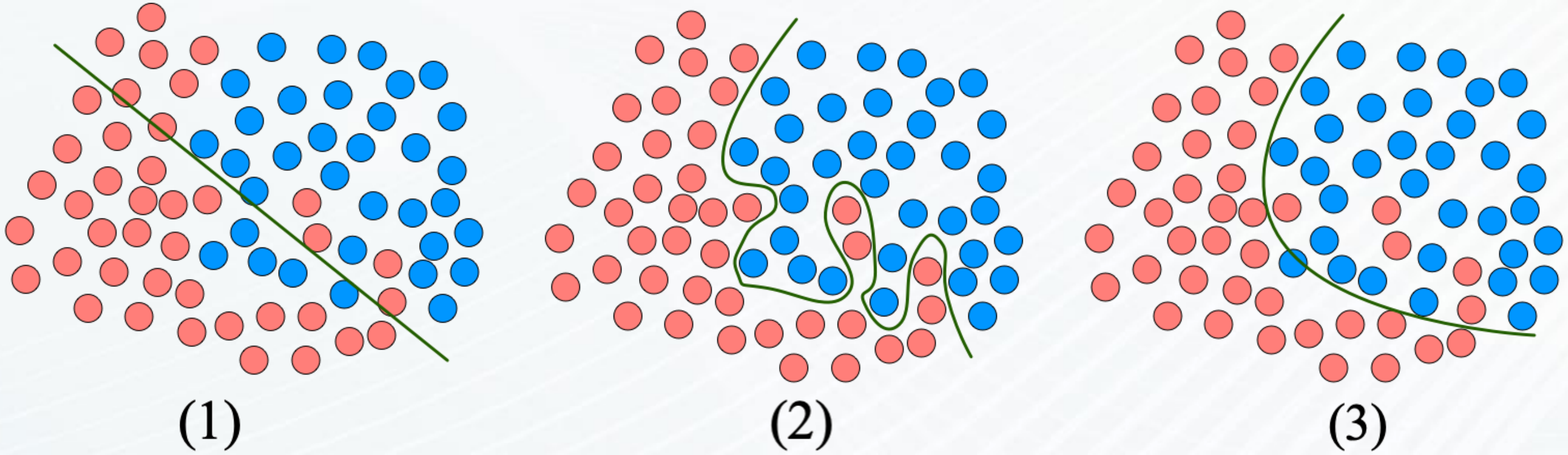
- 模型越简单，经验风险越大，结构风险越小。
- 模型越复杂，经验风险越小，结构风险越大。
- λ用来调和经验风险和结构风险之间的矛盾。



三、模型解释

▶ 过拟合与正则化

- 对于一个通用的分类器模型而言，结构化风险$||\omega||^2$有一个更常用的称谓——正则化项（Regularization term）。
- 奥卡姆剃刀原理（Occam’s Razor）:在所有能够解决当前训练数据集上的分类问题的模型假设中，我们应该选择其中最简单的一个假设。



总 结

软间隔SVM求解

- 问题模型
  - 问题的提出
  - 引入松弛变量
  - 代理损失
  - 软间隔的原始问题
- 问题转化
  - 软间隔的对偶问题
  - 软间隔的KKT条件
  - KKT条件的解释
- 模型解释
  - 经验风险与结构风险
  - 过拟合与正则化

M L & P R

THANK YOU

感谢聆听

MADE BY DONGYUE CHEN