人工智能基础算法第三节
线性有监督分类或回归

于 国 强
清 华 大 学
2025 年 9 月 30 日

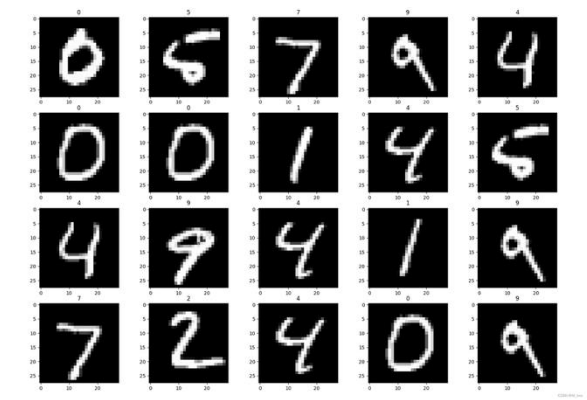
本节课的安排

- 引言
- 线性回归/最小二乘法
- 逻辑回归/罗杰斯特回归
- 偏差– 方差困境/Bias-Variance Dilemma
- 岭回归/Ridge Regression
- LASSO 回归
- 支撑向量机/SVM
- 小结

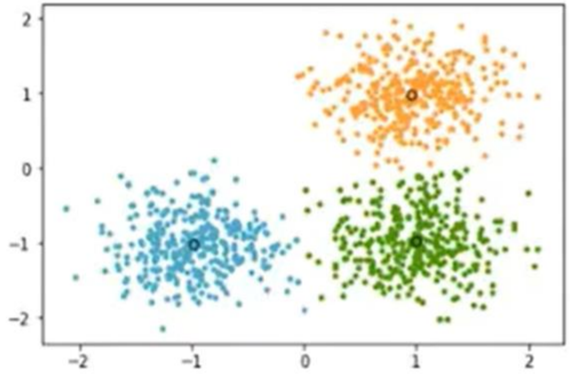
本 节 课 的 安 排

- 引言
- 线性回归/最小二乘法
- 逻辑回归/罗杰斯特回归
- 偏差– 方差困境/Bias-Variance Dilemma
- 岭回归/Ridge Regression
- LASSO 回归
- 支撑向量机/SVM
- 小结

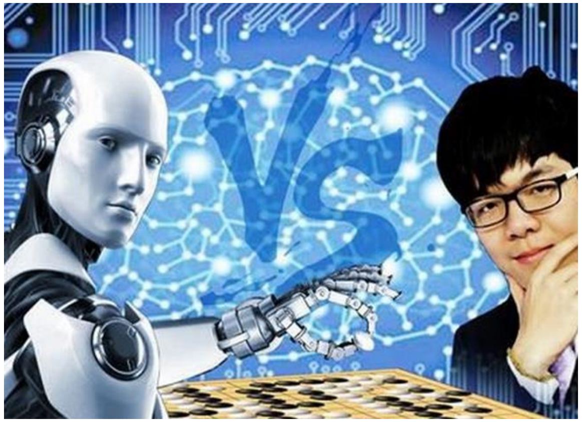
人工智能算法分类



- **有监督学习**
  - 有标签



- 无监督学习
  - 无标签

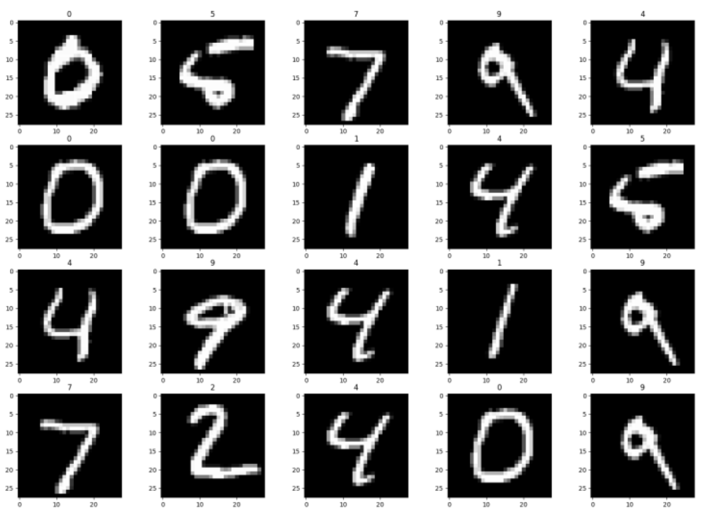


- 强化学习
  - 目标是动态的

- 生成式学习
  - 产生符合给定要求的样本

最近邻分类器

- 最近邻分类器 （Nearest Neighbor Algorithm）
  - 给定一组带有类别标签的样本，对每一个未知标签的样本，找到最近的样本，然后把对应样本的标签赋给该未知样本。
- 例子：　给定图像　和　对应标签



0	5	7	9	4
0	0	1	4	5
4	9	4	1	9
7	2	4	0	9

- 求未知样本  的类别

关于分类、回归和预测

- 分类 vs 回归/预测

  分类：没有可计量的,离散

- 回归 vs 预测

  回归：used more in statistics (old)
  (regression)

  预测：encompasses (回归, classification,
  (prediction)

最近邻分类器优点与缺点

- 优点？

普世

easy to do

- 缺点？

引入线性形式这一约束

- $g(y) = a_0 + B^T\mathrm{x}$
  - y是指待预测量，一般是标量
  - x 是指已知变量或特征，一般是向量
  - a_0是一个标量，需要被估计
  - B是一个向量，也需要被估计
  - $g(y)$是给定的，最简单的$g(y)$形式是 $g(y) \equiv y$
- 线性有监督分类或回归的任务是，
  - 给定训练样本$\{\mathrm{x}_n, y_n\}, n = 1, \ldots, N$, 在$g(y) = a_0 + B^T\mathrm{x}$的假设下，估计/学习未知参数$(a_0, B)$.
- 思考：为什么要引入线性形式这个假设呢？

例子：引入线性形式这一约束

- $g(y) = a_0 + B^T \mathrm{x}$
- 如果y是一连续变量，经常设定$g(y) \equiv \mathrm{y}$，
  - $y = a_0 + B^T \mathrm{x}$
  - 这就是线性回归
- 如果y是二值变量，经常设定$g(y) \equiv \log\left(\frac{\Pr(y==1)}{1-\Pr(y==1)}\right)$
  - $\log\left(\frac{\Pr(y==1)}{1-\Pr(y==1)}\right) = a_0 + B^T \mathrm{x}$
  - 这就是逻辑回归，也叫罗杰斯特回归（Logistic Regression）

本节课的安排

- 引言
- **线性回归/最小二乘法**
- 逻辑回归/罗杰斯特回归
- 偏差– 方差困境/Bias-Variance Dilemma
- 岭回归/Ridge Regression
- LASSO 回归
- 支撑向量机/SVM
- 小结

线性回归/最小二乘法

- 给定训练样本$\{\mathrm{x}_n, y_n\}, n = 1, \dots, N$, 在$\mathrm{y} = a_0 + B^T \mathrm{x}$的假设下，估计/学习未知参数$(a_0, B)$.

- 如何去估计或学习这些未知参数呢?

线性回归/最小二乘法

- 给定训练样本$\{\mathrm{x}_n, y_n\}, n = 1, \ldots, N$, 在$\mathrm{y} = a_0 + B^T\mathrm{x}$的假设下，估计/学习未知参数$(a_0, B)$.

- 优化某一目标函数，最常用之一是优化最小均方误差，
  - $\min_{\boldsymbol{\beta}} \frac{1}{2}\|Y - X\boldsymbol{\beta}\|_2^2 = \min_{\boldsymbol{\beta}} \frac{1}{2}\sum_{n=1}^{N}(y_n - \sum_{m=1}^{M} x_{nm}\beta_m)^2$

线性回归/最小二乘法

- 求解 $\min_{\boldsymbol{\beta}} \frac{1}{2} \|Y - X\boldsymbol{\beta}\|_2^2$

Let's map the intuitive idea to the mathematical symbols:

- $y = a_0 + B^T x$: This is the equation of the line (or hyperplane in multiple dimensions) we are trying to find.
  - a_0 is the **y-intercept**. (Where the line crosses the Y-axis).
  - B is the **slope**. (How steep the line is). $B^T x$ is the matrix way of writing this.
- y_n: The *actual, real* price of the n-th house in your data.
- $\sum x_{nm} \beta_m$: This is the *predicted* price for the n-th house, based on our line. (It's $a_0 + B^T x$ calculated for that specific house).
- $(y_n - \sum x_{nm} \beta_m)$: This is the **error** for a single data point (the vertical distance on the graph).
- $\sum (y_n - \sum x_{nm} \beta_m)^2$: This is the **Sum of Squared Errors** for *all* data points.
- $\min (1/2) \sum (y_n - \sum x_{nm} \beta_m)^2$: The $\min$ means we are on a mission to **find the values of a_0 and B that make this total sum as small as possible.** The $(1/2)$ is often added to make the final math a bit cleaner when we take the derivative, but it doesn't change the "minimum" location.

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

线性回归/最小二乘法

- 优化某一目标函数，最常用之一是优化最小均方误差，
    - $\min_{\boldsymbol{\beta}} \frac{1}{2} \|Y - X\boldsymbol{\beta}\|_2^2 = \min_{\boldsymbol{\beta}} \frac{1}{2} \sum_{n=1}^{N} (y_n - \sum_{m=1}^{M} x_{nm}\beta_m)^2$
- 最小均方误差可以由极大似然估计推导出
    - $y = \beta \mathrm{x} + \epsilon$, where ϵ 是0均值高斯噪声

本节课的安排

- 引言
- 线性回归/最小二乘法
- **逻辑回归/罗杰斯特回归**
- 偏差–方差困境/Bias-Variance Dilemma
- 岭回归/Ridge Regression
- LASSO 回归
- 支撑向量机/SVM
- 小结

逻辑回归/罗杰斯特回归

- 给定训练样本$\{\mathrm{x}_n, y_n\}, n = 1, \dots, N$, 在

$\log\left(\frac{\Pr(y==1)}{1-\Pr(y==1)}\right) = a_0 + B^T\mathrm{x}$的假设下，估计/学习未知参数$(a_0, B)$.

Logistic Regression predicts the Probability of some data point belonging to a category.

- Logistic Regression, 逻辑回归？罗杰斯特回归？

- 因为y是伯努利分布，我们最大化对数似然函数

$$\max_{\boldsymbol{\beta}} \sum_{n=1}^{N} \left[y_n \beta^T \mathrm{x}_n - y_n \log\left(1 + e^{\beta^T \mathrm{x}_n}\right) + (1 - y_n) \log\left(1 + e^{\beta^T \mathrm{x}_n}\right) \right]$$

逻辑回归/罗杰斯特回归

- 我们最大化对数似然函数

$$\max_{\boldsymbol{\beta}} \sum_{n=1}^{N} \left[y_n \beta^T \mathrm{x}_n - y_n \log\left(1 + e^{\beta^T \mathrm{x}_n}\right) + (1 - y_n) \log\left(1 + e^{\beta^T \mathrm{x}_n}\right) \right]$$

- 该最优化问题可以被梯度法或牛顿法求解
  - **本**课程不做此要求
  - **可以利用**现有的程序包

线性回归、逻辑回归潜在的问题

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

引入线性形式这一约束

- $g(y) = a_0 + B^T \mathrm{x}$
  - y是指待预测量，一般是标量
  - x 是指已知变量或特征，一般是向量
  - a_0是一个标量，需要被估计
  - B是一个向量，也需要被估计
  - $g(y)$是给定的，最简单的$g(y)$形式是 $g(y) \equiv y$
- 线性有监督分类或回归的任务是，
  - 给定训练样本$\{\mathrm{x}_n, y_n\}, n = 1, \dots, N$, 在$g(y) = a_0 + B^T \mathrm{x}$的假设下，估计/学习未知参数$(a_0, B)$.
- **思考：为什么要引入线性形式这个假设呢?**

本节课的安排

- 引言
- 线性回归/最小二乘法
- 逻辑回归/罗杰斯特回归
- **偏差– 方差困境/Bias-Variance Dilemma**
- 岭回归/Ridge Regression
- LASSO 回归
- 支撑向量机/SVM
- 总结

Bias Variance Dilemma

- Prediction Error
  - Assume $y = f(x) + \epsilon, E(\epsilon) = 0,$ and $var(\epsilon) = \sigma^2$, then the prediction error of the estimate $\hat{f}(x)$ is

$$Error_{\hat{f}(x)} = E\left[\left(y - \hat{f}(x)\right)^2\right]$$

$$= \sigma^2 + \left[E\hat{f}(x) - f(x)\right]^2 + E\left[\left(\hat{f}(x) - E\hat{f}(x)\right)^2\right]$$

$$= \sigma^2 + bias^2\left(\hat{f}(x)\right) + var\left(\hat{f}(x)\right)$$

  - OLS estimates often have low bias but large variance

Introducing constraint (bias)

- Hope that the introduction of a small bias will substantially reduce the variance

$$Error_{\hat{f}(x)} = \sigma^2 + bias^2\left(\hat{f}(x)\right) + var\left(\hat{f}(x)\right)$$

- Penalty ≈ constraint ≈ bias ≈ the prior knowledge

- 你能想到什么样的bias呢?

本节课的安排

- 引言
- 线性回归/最小二乘法
- 逻辑回归/罗杰斯特回归
- 偏差–方差困境/Bias-Variance Dilemma
- **岭回归/Ridge Regression**
- LASSO 回归
- 支撑向量机/SVM
- 小结

Ridge constraint (L2 norm)

$$\min_{\boldsymbol{\beta}} \frac{1}{2} \sum_{n=1}^{N} \left(y_n - \sum_{m=1}^{M} x_{nm} \beta_m \right)^2$$

s.t. $\sum_{m=1}^{M} \beta_m^2 \leq t$

Ridge constraint 的 等 价 形 式

$$\min_{\boldsymbol{\beta}} \frac{1}{2} \sum_{n=1}^{N} \left(y_n - \sum_{m=1}^{M} x_{nm} \beta_m \right)^2 + \lambda \sum_{m=1}^{M} \beta_m^2$$

岭回归 / Ridge Regression

$$\min_{\boldsymbol{\beta}} \frac{1}{2} \sum_{n=1}^{N} \left(y_n - \sum_{m=1}^{M} x_{nm} \beta_m \right)^2 + \lambda \sum_{m=1}^{M} \beta_m^2$$

- **作业**：写成矩阵形式，推导出Ridge Regression的解。

本节课的安排

- 引言
- 线性回归/最小二乘法
- 逻辑回归/罗杰斯特回归
- 偏差– 方差困境/Bias-Variance Dilemma
- 岭回归/Ridge Regression
- **LASSO 回归**
- 支撑向量机/SVM
- 小结

LASSO 约束

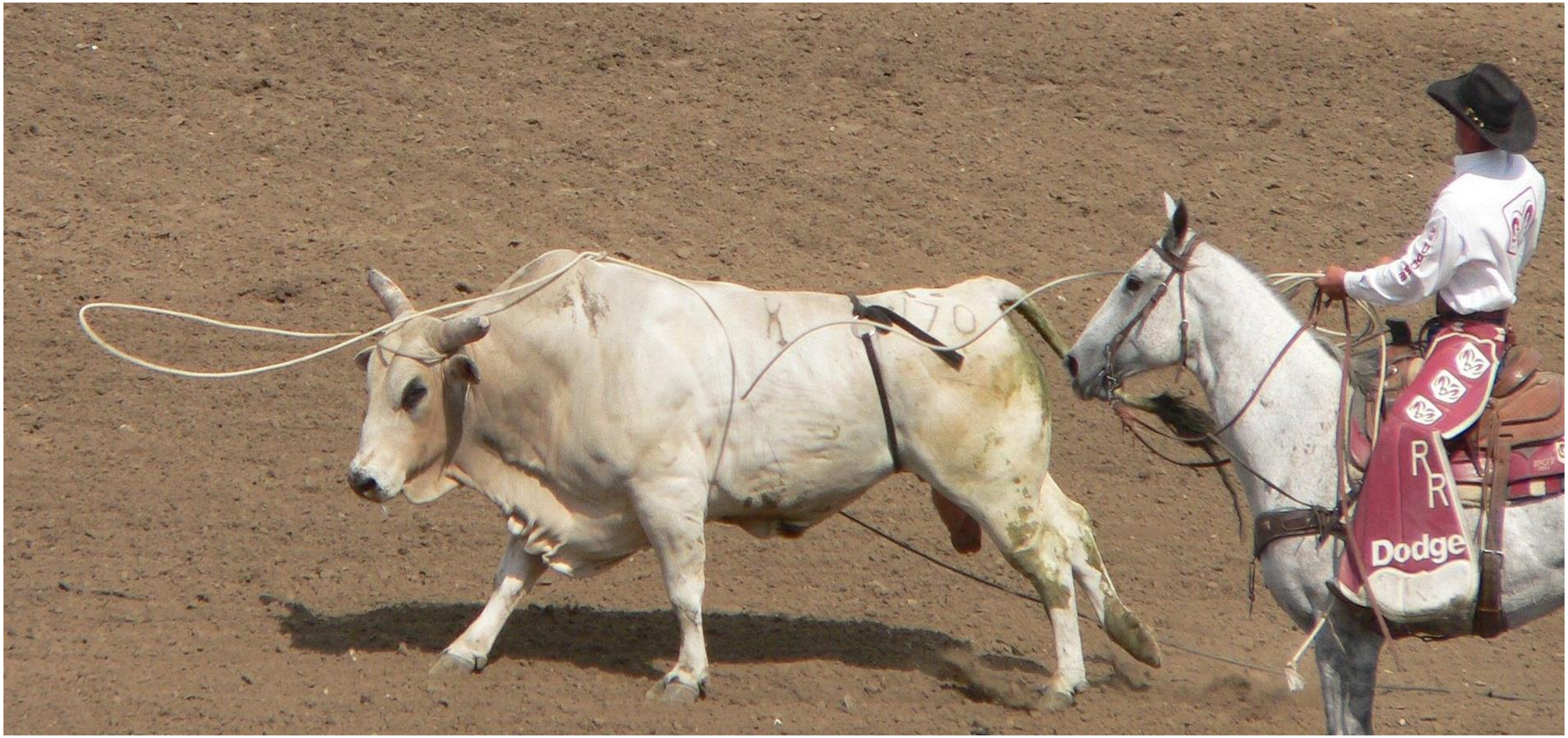
$$\min_{\boldsymbol{\beta}} \frac{1}{2} \sum_{n=1}^{N} \left(y_n - \sum_{m=1}^{M} x_{nm} \beta_m \right)^2$$

s.t. $\sum_{m=1}^{M} |\beta_m| \leq t$

LASSO 回 归

- **L**east **A**bsolute **S**hrinkage and **S**election **O**perator

- A loop of rope designed as a restraint to be thrown around a target and tightened when pulled.



LASSO 约束的等价形式

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\sum_{n=1}^{N}\left(y_n - \sum_{m=1}^{M} x_{nm}\beta_m\right)^2$$

s.t. $\sum_{m=1}^{M}|\beta_m| \leq t$

===

$$\min_{\boldsymbol{\beta}} \left\{\frac{1}{2}\sum_{n=1}^{N}\left(y_n - \sum_{m=1}^{M} x_{nm}\beta_m\right)^2 + \gamma \sum_{m=1}^{M}|\beta_m|\right\}$$

LASSO 回归的求解

$$\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} \sum_{n=1}^{N} \left(y_n - \sum_{m=1}^{M} x_{nm} \beta_m \right)^2 + \gamma \sum_{m=1}^{M} |\beta_m| \right\}$$

凸函数，而且形式特殊，可以一个变量一个变量的迭代求解。

L0 Norm 约束

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\sum_{n=1}^{N}\left(y_n - \sum_{m=1}^{M} x_{nm}\beta_m\right)^2$$

s.t. $\quad \sum_{m=1}^{M}|\beta_m|^0 \leq t$

Note:

$|\beta_m|^0 = 1$ if $\beta_m \neq 0$;

$|\beta_m|^0 = 0$ if $\beta_m = 0$.

哪种约束最好？

Bias-variance dilemma

$$\min_{\boldsymbol{\beta}} \frac{1}{2} \sum_{n=1}^{N} \left(y_n - \sum_{m=1}^{M} x_{nm} \beta_m \right)^2$$

s.t. $\sum_{m=1}^{M} |\beta_m|^0 \leq t$?

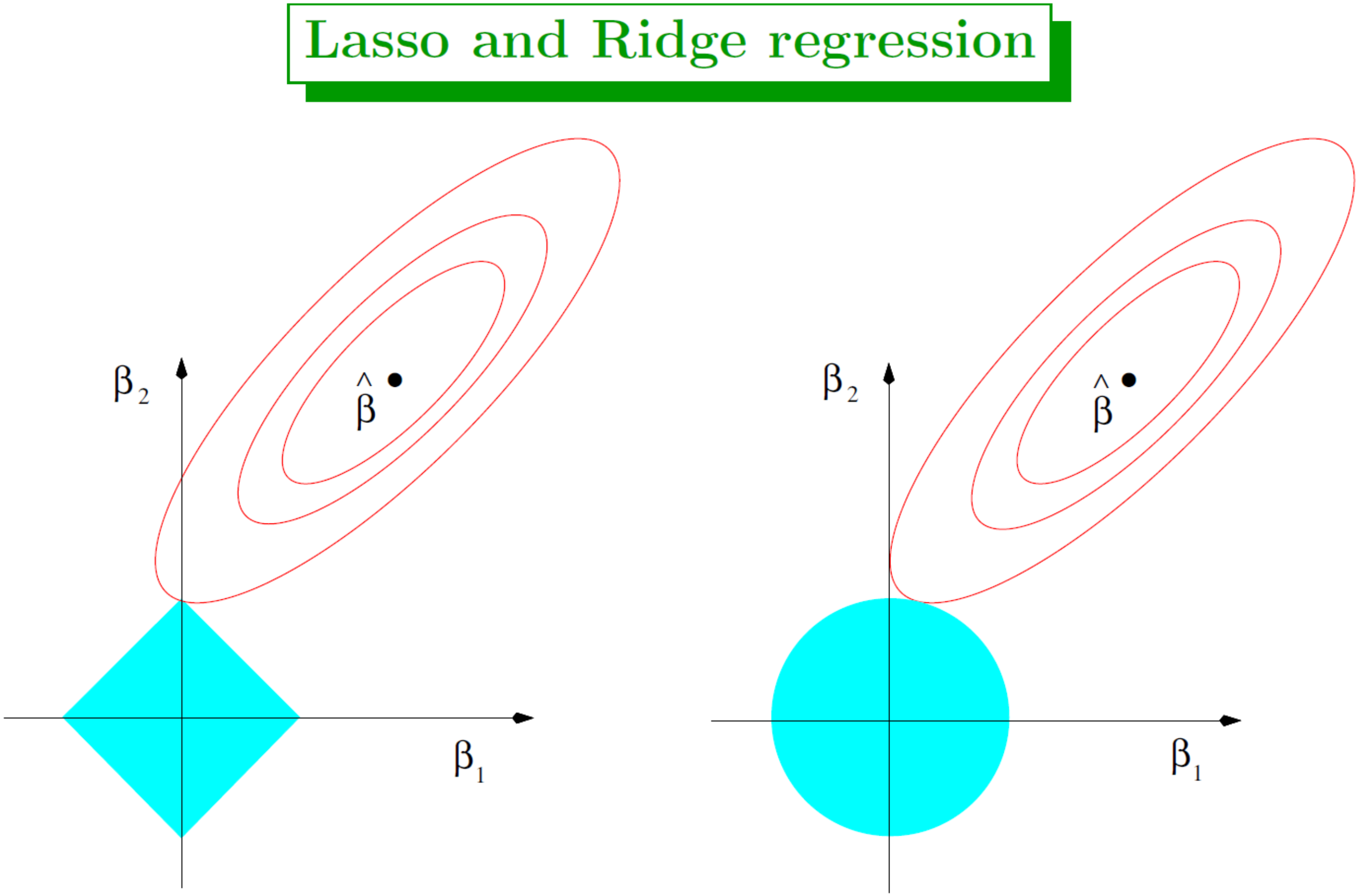
$\sum_{m=1}^{M} |\beta_m|^1 \leq t$?

$\sum_{m=1}^{M} |\beta_m|^2 \leq t$?

LASSO 和 Ridge 回归的差异

- Lasso will force some parameters to be zero, hence a sparse solution, which can be easily interpreted.
- While ridge simply reduces the magnitude by a factor.

LASSO 和 Ridge 的几何解释



L1 constraint:

$|\beta_1| + |\beta_2| \leq 1$

L2 constraint:

$\beta_1^2 + \beta_2^2 \leq 1$

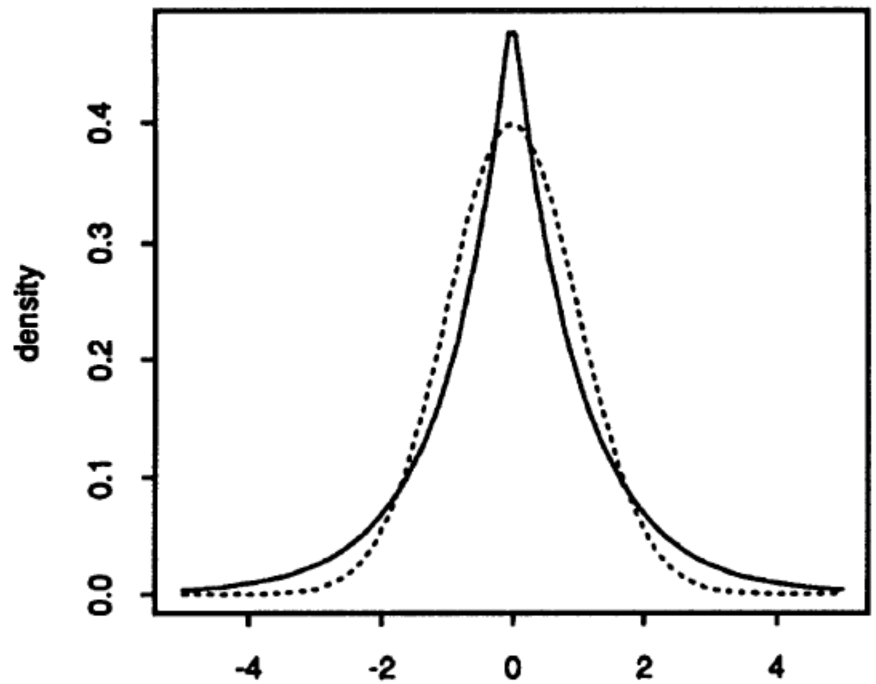
LASSO 和 Ridge 的概率解释

Lasso as Bayes Estimate

Assume that y follows a Gaussian distribution $G(X\beta, \sigma^2)$, and β has the Laplacian prior distribution as:

$$f(\beta_j) = \frac{1}{2\tau} \exp\left(-\frac{|\beta_j|}{\tau} \right)$$

Then, we can derive the lasso regression estimate as the **Bayes posterior mode**.



Similarly, ridge form can be derived by assuming β has a Gaussian prior distribution.

哪种约束最好？

Bias-variance dilemma

$$\min_{\boldsymbol{\beta}} \frac{1}{2} \sum_{n=1}^{N} \left(y_n - \sum_{m=1}^{M} x_{nm} \beta_m \right)^2$$

s.t. $\sum_{m=1}^{M} |\beta_m|^0 \leq t$

$\sum_{m=1}^{M} |\beta_m|^1 \leq t$

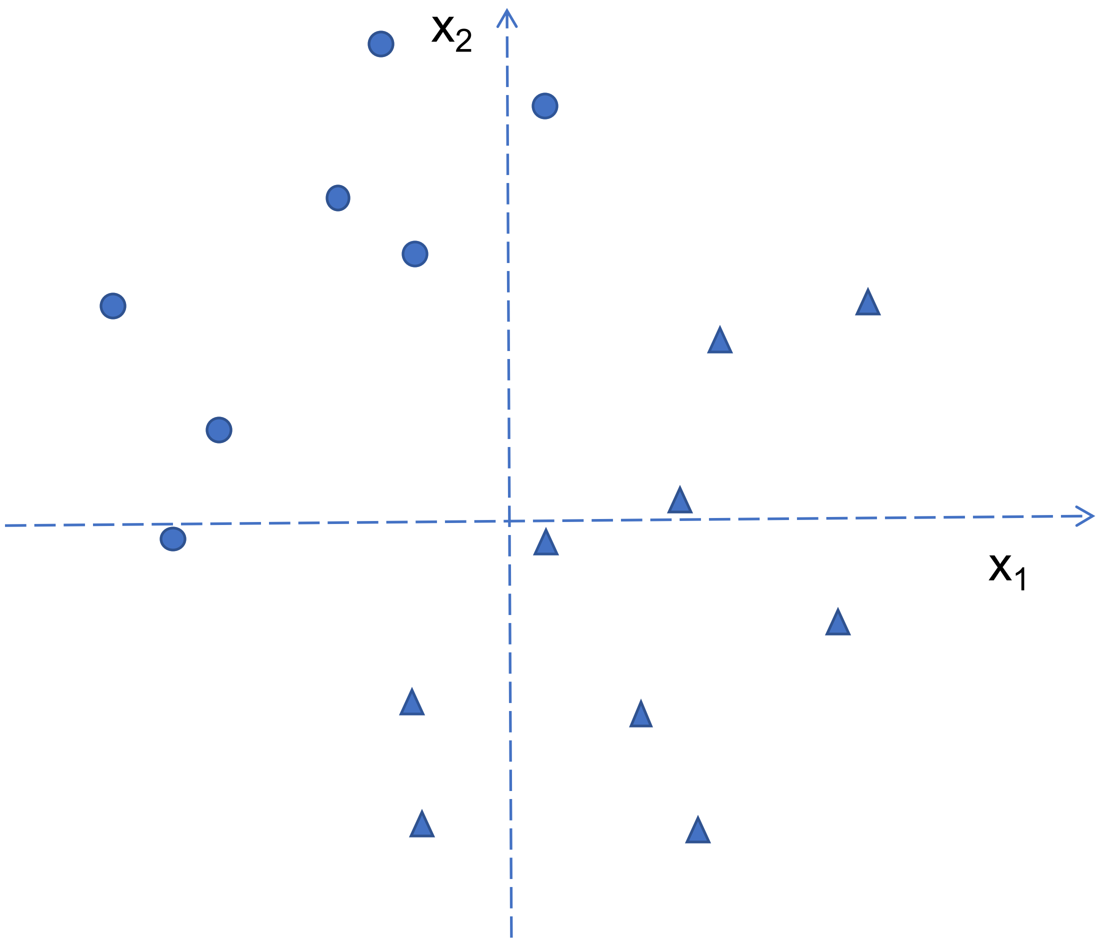
$\sum_{m=1}^{M} |\beta_m|^2 \leq t$

本节课的安排

- 引言
- 线性回归/最小二乘法
- 逻辑回归/罗杰斯特回归
- 偏差– 方差困境/Bias-Variance Dilemma
- 岭回归/Ridge Regression
- LASSO 回归
- **支撑向量机/SVM**
- 小结

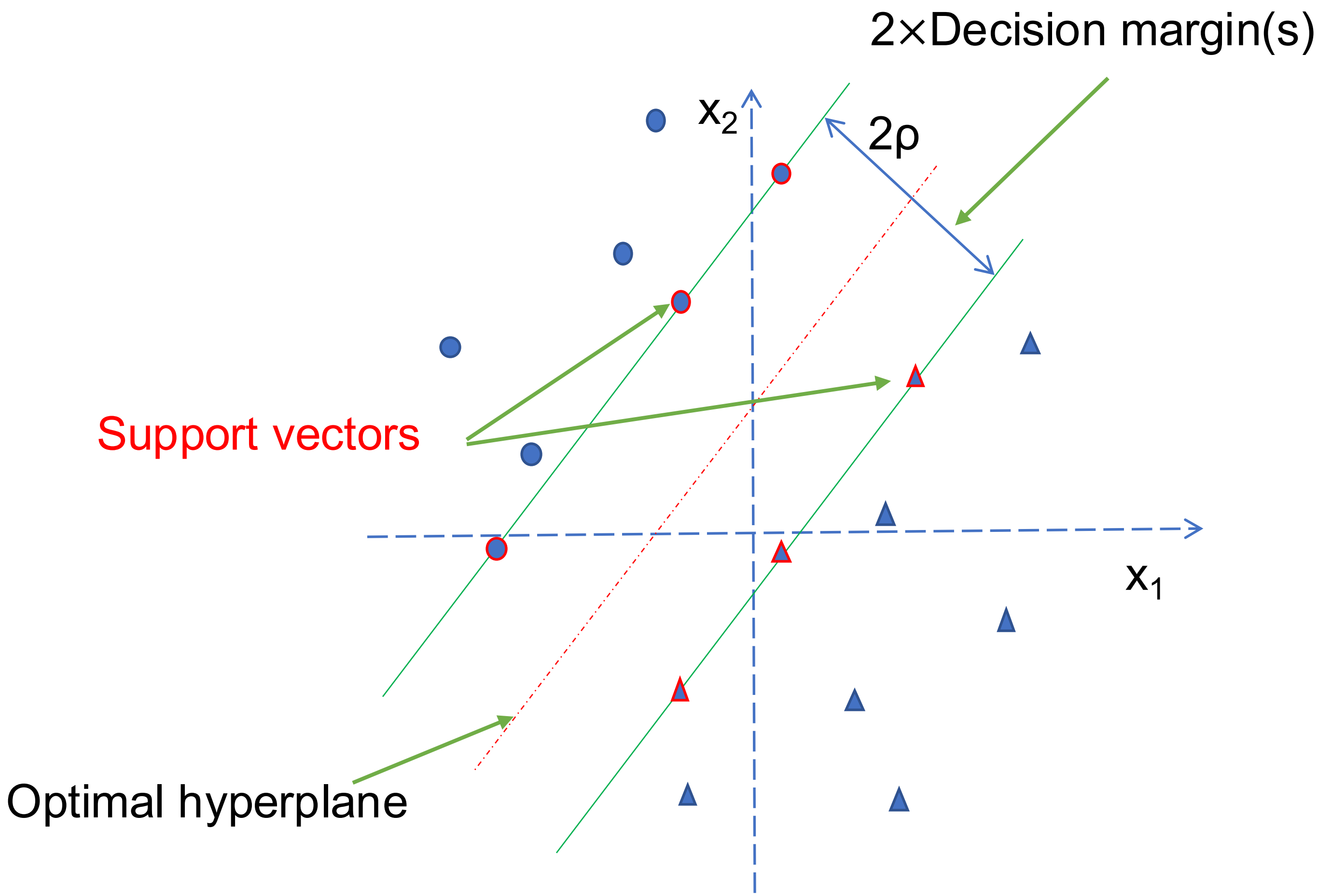
支撑向量机的思想

- For two classes that are separable, there are an infinite number of hyperplanes that perfectly separate the two classes in the training set.



支撑向量机的思想

- SVM hypothesizes that the linear function that **maximizes the margin** is the best one for future samples.
  - Margin for any given linear function $g(x)$ is defined as the least distance between training samples and the hyperplane $g(x) = 0$.

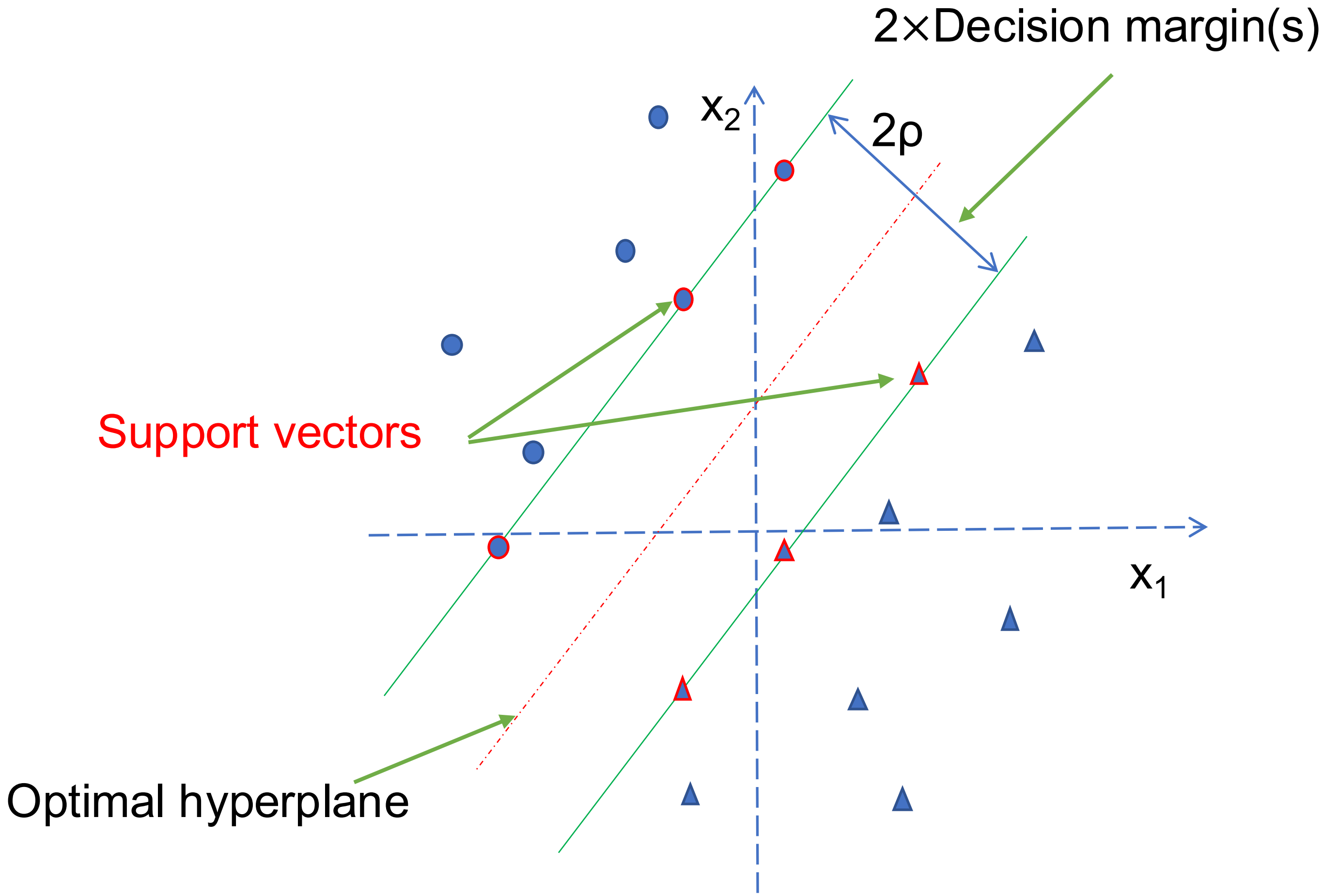


支撑向量机可以被描述成一个优化问题

- Given the training sample $\{(\mathbf{x}_i, d_i)\}_i^N$, find optimal $\mathbf{w}_0$ and b_0 such that

$$\{\mathbf{w}_0, b_0\} = \operatorname{argmin}\left\{\Phi(\mathbf{w}, b) = \frac{1}{2}\mathbf{w}^T\mathbf{w}\right\},$$

$$\mathbf{s.t.} \quad d_i(\mathbf{w}^T\mathbf{x} + b) \geq 1$$



SVM: linear non-separable pattern

- Thus, the overall objective function becomes

$$\Phi(\mathbf{w},\boldsymbol{\xi}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}\xi_i$$

where the parameter C controls the tradeoff between model complexity and the number of non-separable points.

- The primal problem:

Given the training sample $\{(\mathbf{x}_i, d_i)\}_{i=1}^{N}$, find optimum $\mathbf{w}$, b, and $\boldsymbol{\xi}$ that minimize the cost function

$$\Phi(\mathbf{w},\boldsymbol{\xi}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}\xi_i$$

$$\text{subject to } \begin{cases} d_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 1 - \xi_i, \\ \xi_i \geq 0. \end{cases}$$

求解支撑向量机

- Thus, the overall objective function becomes

$$\Phi(\mathbf{w},\xi) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}\xi_i$$

where the parameter C controls the tradeoff between model complexity and the number of non-separable points.

- The primal problem:

Given the training sample $\{(\mathbf{x}_i, d_i)\}_{i=1}^{N}$, find optimum $\mathbf{w}$, b, and ξ that minimize the cost function

$$\Phi(\mathbf{w},\xi) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}\xi_i$$

$$\text{subject to} \begin{cases} d_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 1 - \xi_i, \\ \xi_i \geq 0. \end{cases}$$

- 这是一个凸优化问题。
- 需要**比**较强的凸优化知识。

各种模型的异同

- 线性回归/最小二乘法

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|Y - X\boldsymbol{\beta}\|_2^2 = \min_{\boldsymbol{\beta}} \frac{1}{2}\sum_{n=1}^{N}\left(y_n - \sum_{m=1}^{M} x_{nm}\beta_m\right)^2$$

- 岭回归/Ridge Regression

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\sum_{n=1}^{N}\left(y_n - \sum_{m=1}^{M} x_{nm}\beta_m\right)^2 + \lambda \sum_{m=1}^{M} \beta_m^2$$

- LASSO 回归

$$\min_{\boldsymbol{\beta}} \left\{\frac{1}{2}\sum_{n=1}^{N}\left(y_n - \sum_{m=1}^{M} x_{nm}\beta_m\right)^2 + \gamma \sum_{m=1}^{M} |\beta_m|\right\}$$

- 支撑向量机/SVM

$$\min_{w,b,\zeta} \left\{\sum_{m=1}^{M} w_m^2 + C\sum_{n=1}^{N} \zeta_n\right\}$$

$$\text{s.t. } d_n(w^T x_n + b) \geq 1 - \zeta \ \&\ \zeta \geq 0$$

各种模型的异同

- 线性回归/最小二乘法

$$\min_{\boldsymbol{\beta}} \frac{1}{2} \|Y - X\boldsymbol{\beta}\|_2^2 = \min_{\boldsymbol{\beta}} \frac{1}{2} \sum_{n=1}^{N} \left(y_n - \sum_{m=1}^{M} x_{nm}\beta_m \right)^2$$

- 岭回归/Ridge Regression

$$\min_{\boldsymbol{\beta}} \frac{1}{2} \sum_{n=1}^{N} \left(y_n - \sum_{m=1}^{M} x_{nm}\beta_m \right)^2 + \lambda \sum_{m=1}^{M} \beta_m^2$$

- LASSO 回归

$$\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} \sum_{n=1}^{N} \left(y_n - \sum_{m=1}^{M} x_{nm}\beta_m \right)^2 + \gamma \sum_{m=1}^{M} |\beta_m| \right\}$$

- 支撑向量机/SVM的等价形式

$$\min_{w,b,\zeta} \left\{ \sum_{m=1}^{M} w_m^2 + C \sum_{n=1}^{N} \max\left(0, 1 - d_n\left(w^T \mathrm{x}_n + b\right)\right) \right\}$$

如何寻找各种方法的最优参数？

- $\min_{\boldsymbol{\beta}} \frac{1}{2}\sum_{n=1}^{N}(y_n - \sum_{m=1}^{M} x_{nm}\beta_m)^2 + \lambda \sum_{m=1}^{M} \beta_m^2$

- $\min_{\boldsymbol{\beta}} \left\{\frac{1}{2}\sum_{n=1}^{N}(y_n - \sum_{m=1}^{M} x_{nm}\beta_m)^2 + \gamma \sum_{m=1}^{M} |\beta_m|\right\}$

- $\min_{w,b,\zeta} \{\sum_{m=1}^{M} w_m^2 + C \sum_{n=1}^{N} \zeta_n\}$

本节课的小结

- 优化技术
- 偏差– 方差困境/Bias-Variance Dilemma
- 交叉检验寻找最优参数
- 线性回归， 逻辑回归， 岭回归， LASSO回归，支撑向量机 等具体模型

下一节课内容

- 人工神经网络 / 深度神经网络

Thank you!