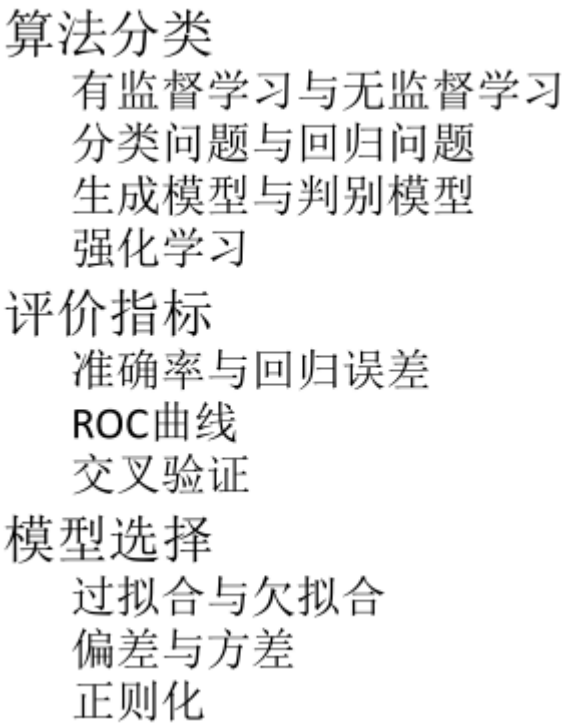
4 基本概念

4.1 本集介绍

这节课介绍机器学习基本的概念，所学内容如下图。



- 算法分类
- 评价指标
- 模型选择

老师提到一个泛化的概念。

4.2 监督信号

- 监督信号

  小孩学习东西的例子。

  半监督学习是可以归类到有监督学习里面的。

4.3 有监督学习

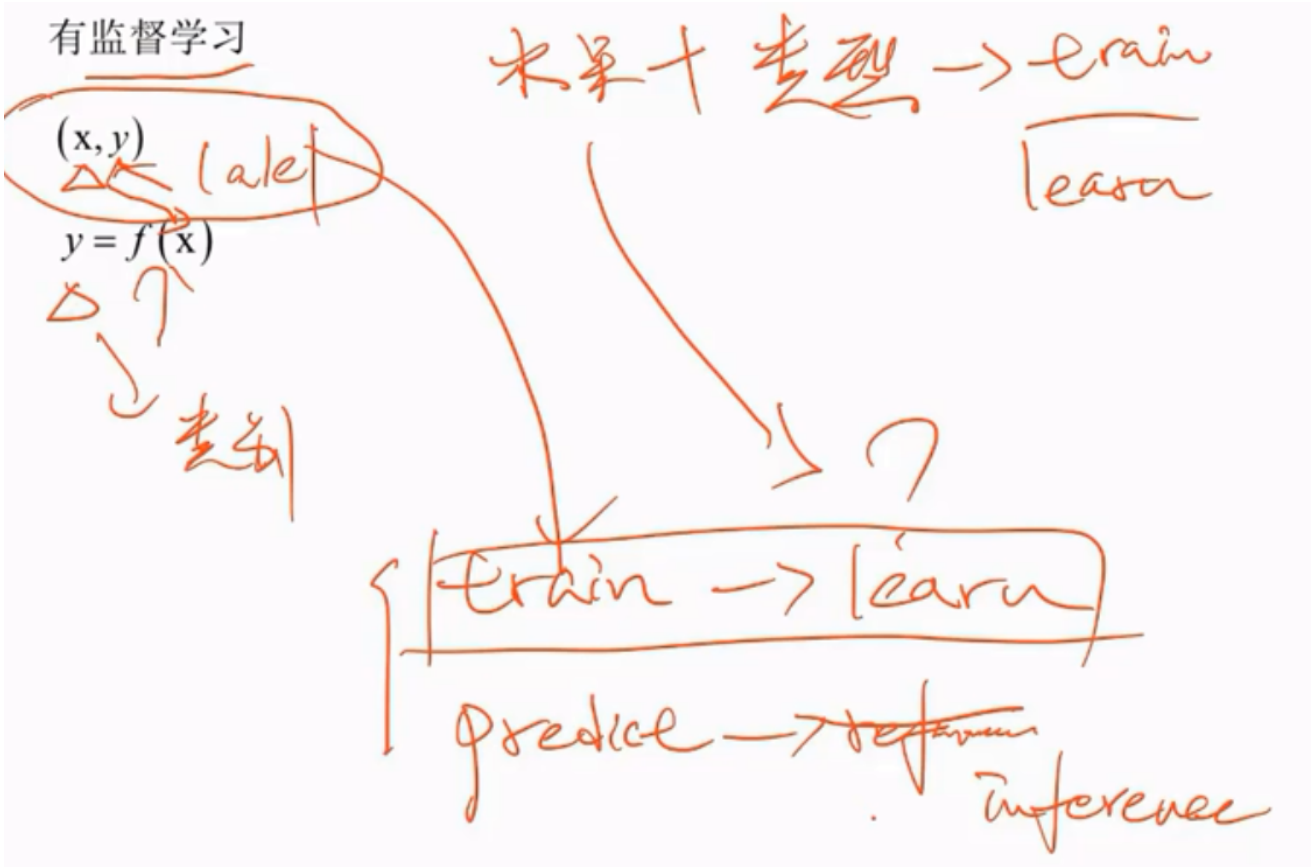
分为学习和预测两个过程。

有监督学习

$$(x, y)$$

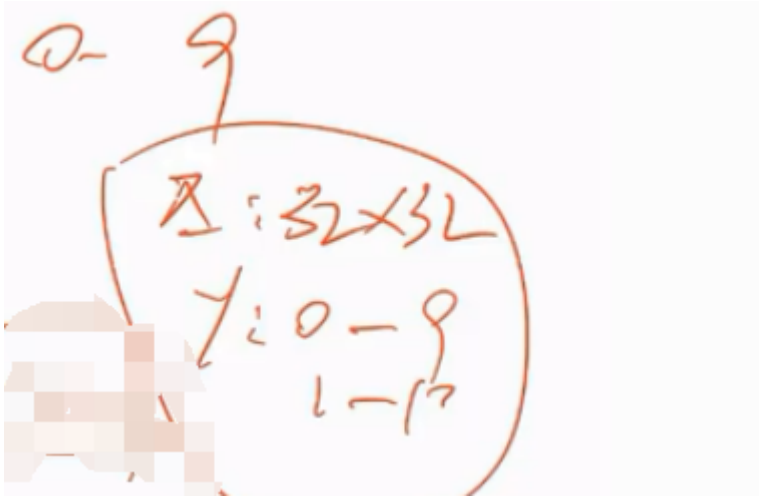
$$y = f(x)$$

预测的例子



有监督学习，处理特征标签值。映射函数。

例子：给你一张图像，预测是0到9中的哪个数字，输入图片的特征向量，输出类别标签



4.4 无监督学习

无监督学习
聚类
数据降维

可以认为像自学一样。

这个概念在吴恩达的课程里的谷歌新闻分类最经典。

一个网页，有政治，经济，军事登录类别。

维数灾难：维数太高，人无法理解，机器学习算法处理起来也困难。

一般人最高能直观理解到三维空间

4.5 强化学习

强化学习
根据环境数据预测动作
最大化奖励值

强化学习源于行为心理学。

- 自动驾驶

4.6 分类问题

分类问题属于监督学习的一部分。

$$\text{分类问题}$$

$$\mathbb{R}^n \to \mathbb{Z}$$

$$\operatorname{sgn}\left(\mathrm{w}^{\mathrm{T}}\mathrm{x}+b\right)$$

n维向量到整数的预测。

有一种特殊的分类问题，叫做二分类问题。如人脸检测问题是属于二分类问题。

4.7 回归问题

$$\text{回归问题}$$

$$\mathbb{R}^n \to \mathbb{R}$$

$$f\left(\mathrm{x}\right)=\mathrm{w}^{\mathrm{T}}\mathrm{x}+b \qquad L=\frac{1}{2l}\sum_{i=1}^{l}\left(f\left(\mathrm{x}_i\right)-y_i\right)^2$$

4.8 线性回归

线性回归

$f(\mathrm{x}) = \mathrm{w}^{\mathrm{T}}\mathrm{x} + b \qquad w_1x_1 + w_2x_2 + \ldots + w_nx_n + b$

(x_i, y_i)

$[\mathrm{w}, b] \to \mathrm{w}$
$[\mathrm{x}, 1] \to \mathrm{x}$

$$L = \frac{1}{2l}\sum_{k=1}^{l}\left(\mathrm{w}^{\mathrm{T}}\mathrm{x}_k - y_k\right)^2 \qquad \frac{\partial L}{\partial w_i} = \frac{1}{l}\sum_{k=1}^{l}\left(\mathrm{w}^{\mathrm{T}}\mathrm{x}_k - y_k\right)x_{ki}$$

$$\frac{\partial^2 L}{\partial w_i \partial w_j} = \frac{1}{l}\sum_{k=1}^{l}x_{ki}\frac{\partial\left(\mathrm{w}^{\mathrm{T}}\mathrm{x}_k - y_k\right)}{\partial w_j} = \frac{1}{l}\sum_{k=1}^{l}x_{ki}x_{kj}$$

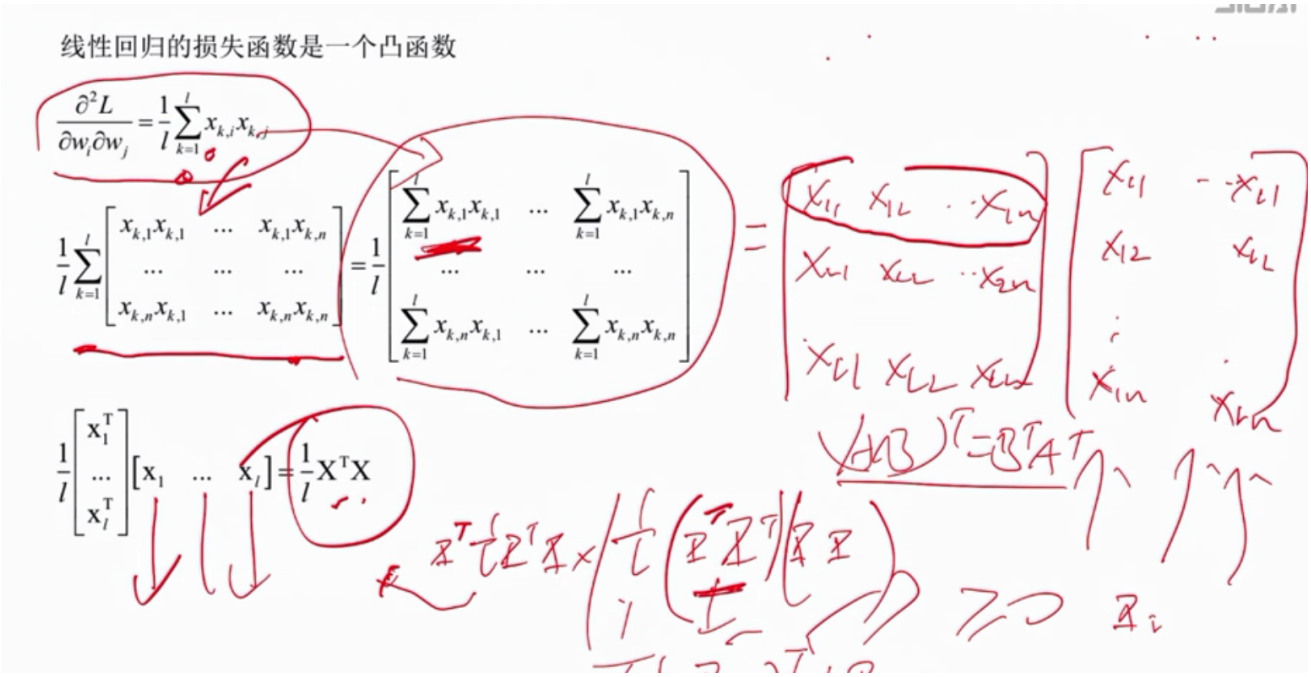
线性回归的损失函数是一个凸函数

$$\frac{\partial^2 L}{\partial w_i \partial w_j} = \frac{1}{l}\sum_{k=1}^{l}x_{k,i}x_{k,j}$$

$$\frac{1}{l}\sum_{k=1}^{l}\begin{bmatrix} x_{k,1}x_{k,1} & \cdots & x_{k,1}x_{k,n} \\ \cdots & \cdots & \cdots \\ x_{k,n}x_{k,1} & \cdots & x_{k,n}x_{k,n} \end{bmatrix} = \frac{1}{l}\begin{bmatrix} \sum_{k=1}^{l}x_{k,1}x_{k,1} & \cdots & \sum_{k=1}^{l}x_{k,1}x_{k,n} \\ \cdots & \cdots & \cdots \\ \sum_{k=1}^{l}x_{k,n}x_{k,1} & \cdots & \sum_{k=1}^{l}x_{k,n}x_{k,n} \end{bmatrix}$$

$$\frac{1}{l}\begin{bmatrix} \mathrm{x}_1^{\mathrm{T}} \\ \cdots \\ \mathrm{x}_l^{\mathrm{T}} \end{bmatrix}\begin{bmatrix} \mathrm{x}_1 & \cdots & \mathrm{x}_l \end{bmatrix} = \frac{1}{l}\mathrm{X}^{\mathrm{T}}\mathrm{X}$$

具体的细节如下图



4.9 判别模型与生成模型

判别模型

$$y = f(\mathrm{x}) \qquad p(y|\mathrm{x})$$

生成模型

生成模型

$$p(\mathrm{x},y)=p(\mathrm{x}|y)p(y)$$

另外一种定义：

$$p(\mathrm{x}|y)$$

$$p(y|\mathrm{x})$$

生成模型：贝叶斯分类器，高斯混合模型，隐马尔可夫模型，受限玻尔兹曼机，生成对抗网络等
判别模型：决策树，kNN算法，人工神经网络，支持向量机，logistic回归，AdaBoost算法等

与之相对应的是生成模型

生成模型：贝叶斯分类器，高斯混合模型，隐没尔科夫模型，受限玻尔兹曼机，生成对抗网络等

判别模型：决策树，KNN算法，人工神经网络，支持向量机，逻辑回归，AdaBoost算法等。

4.10 准确率

准确率是算法的评估指标。

准确率　　训练集
　　测试集

$$\frac{\text{正确分类的样本数}}{\text{测试样本总数}}$$

同一类问题，不同的算法都快可以解决。

分类问题的准确率，正确分类的样本数除以测试样本总数。

具有训练集和测试。

4.11 回归误差

$$\frac{1}{l}\sum_{i=1}^{l}\left(f\left(\mathrm{x}_i\right)-y_i\right)^2$$

预测值和真实值的差的平方

4.12 精度与召回率

精度
召回率

$$\mathrm{TP}/\left(\mathrm{TP}+\mathrm{FP}\right)$$
$$\mathrm{TP}/\left(\mathrm{TP}+\mathrm{FN}\right)$$

	判定为正	判定为负
正样本	TP	FN
负样本	FP	TN

前面说过二分类问题， 医生判别病人是否得病。

定义

- 精度：
- 召回率：这个指标是越大越大越好的。
- TP: 真阳性，真正的正样本数，
- FP：假阳性，错误的正样本数
- TN：真阴性
- FN：假阴性

判定为正样本

$$\mathrm{TP}/\left(\mathrm{TP}+\mathrm{FP}\right)$$

正样本数量

正样本	TP	FN

$$TP/(TP+FN)$$

4.13 ROC曲线

ROC曲线
真阳率（检测率）
假阳率（误报率）

$$TPR = TP/(TP+FN)$$
$$FPR = FP/(FP+TN)$$

$$\mathrm{sgn}(f(x))$$

$$\mathrm{sgn}(f(x)+\xi)$$

灵敏度

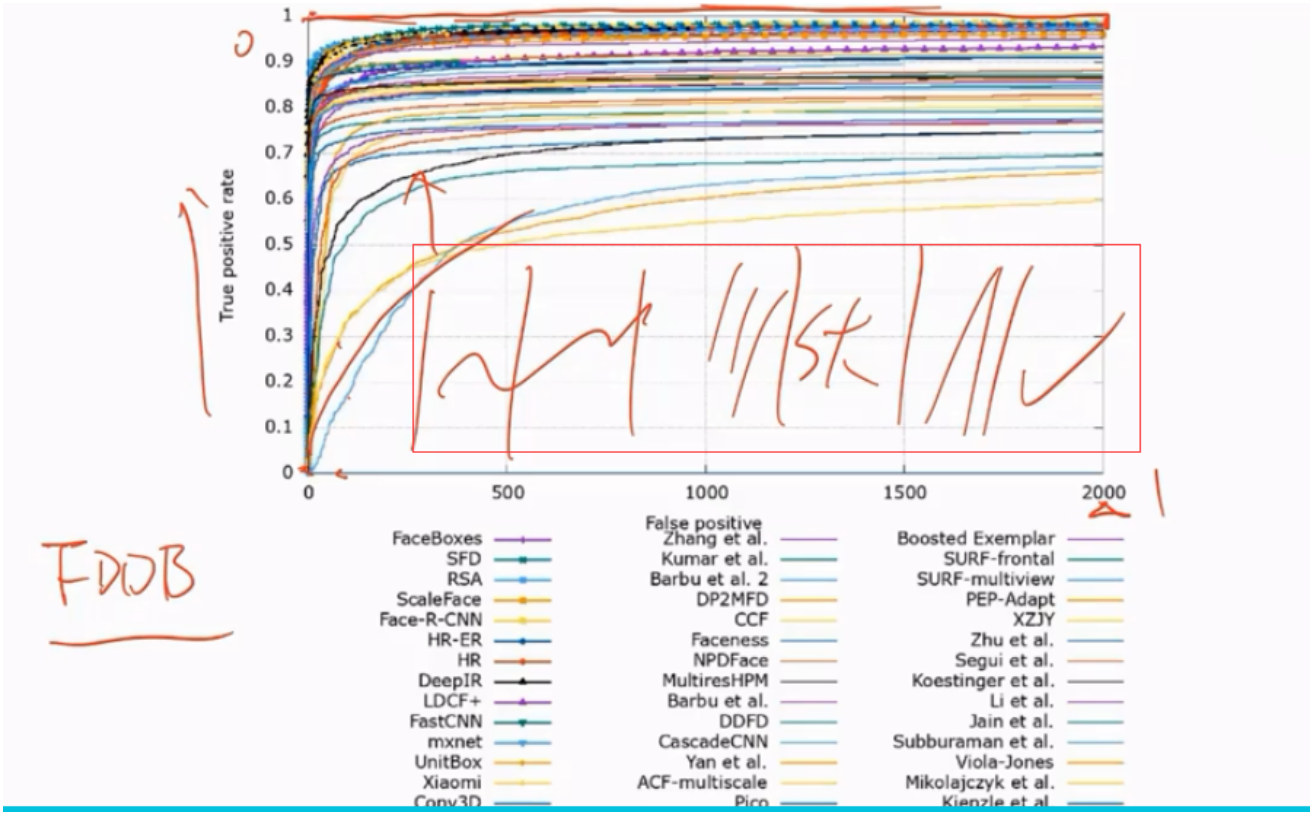
接收机操作曲线 ROC

- 真阳率（检测率） TPR=TP/(TP+FN) 做纵轴
- 假阳率（误报率） FPR=FP/(FP+TN) 做横轴

检测率随着增函数

横坐标是误报率，纵坐标是检测率。

ROC曲线下面的面积越大越好，越在上面的曲线越好。



4.14 混淆矩阵

说完二分类问题的评价指标

我们来说多分类问题的评价指标；混淆矩阵。

$$\begin{bmatrix} c_{11} & \cdots & c_{1k} \\ \cdots & \cdots & \cdots \\ c_{k1} & \cdots & c_{kk} \end{bmatrix}$$

每一行是样本数，每一列代表不同的类。

主对角线上为正确的分类。

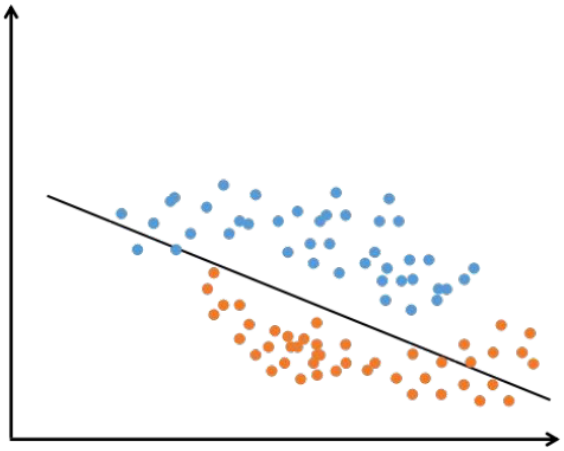
4.15 交叉验证

CV (cross validation)

把样本划分成训练集和测试集。它们不相交。

4.16 欠拟合

欠拟合
训练得到的模型在训练集上表现差，没有学到数据的规律。引起欠拟合的原因有模型本身过于简单，例如数据本身是非线性的但使用了线性模型；特征数太少无法正确的建立映射关系。



图中例子为线性不可分的问题。

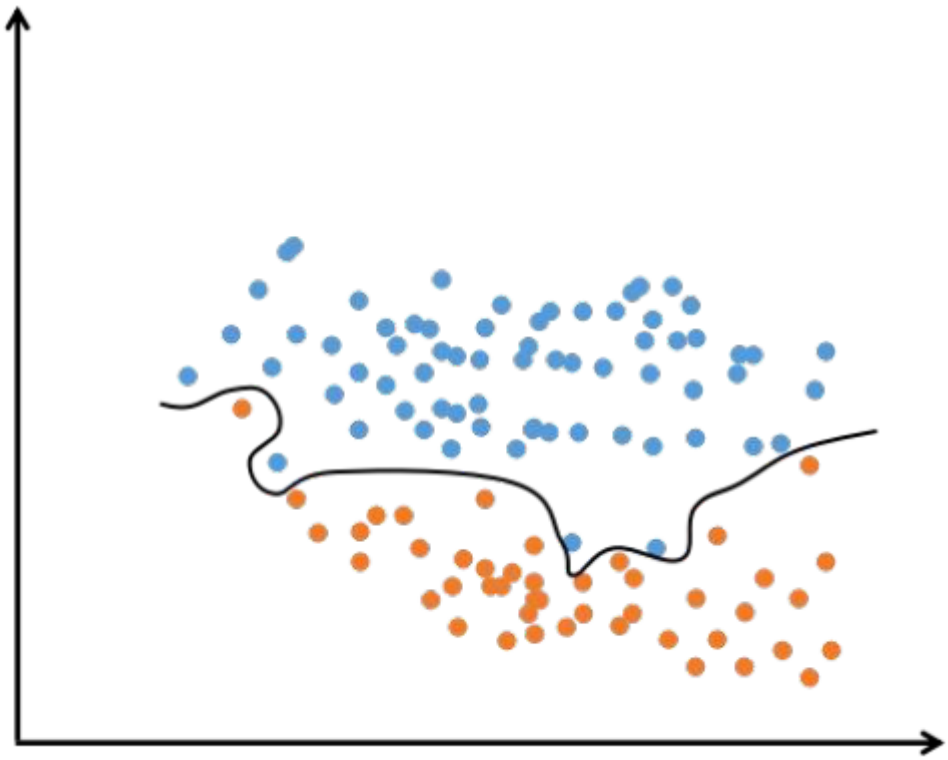
4.17 过拟合

过拟合

直观表现是在训练集上表现好，但在测试集上表现不好，推广泛化性能差。过拟合产生的根本原因是训练数据包含抽样误差，在训练时模型将抽样误差也进行了拟合。所谓抽样误差，是指抽样得到的样本集和整体数据集之间的偏差。

引起过拟合的可能原因有：

模型本身过于复杂，拟合了训练样本集中的噪声。此时需要选用更简单的模型，或者对模型进行裁剪。训练样本太少或者缺乏代表性。此时需要增加样本数，或者增加样本的多样性。训练样本噪声的干扰，导致模型拟合了这些噪声，这时需要剔除噪声数据或者改用对噪声不敏感的模型。



- 在训练集上表现好，在测试集上表现不好。
- 模型本身过于复杂，拟合了训练样本集中的噪声。

4.18 欠拟合和过拟合的总结

怎么判断？

训练集上的表现	测试集上的表现	结论
不好	不好	欠拟合
好	不好	过拟合
好	好	适度拟合

4.19 偏差与方差分解

前面说泛化

偏差 $\text{Bias}\left(\hat{f}(x)\right)=E\left(\hat{f}(x)-f(x)\right)$

方差 $\text{Var}\left(\hat{f}(x)\right)=E\left(\hat{f}^2(x)\right)-E^2\left(\hat{f}(x)\right)$

$$E\left(\left(y-\hat{f}(x)\right)^2\right)=\text{Bias}^2\left(\hat{f}(x)\right)+\text{Var}\left(\hat{f}(x)\right)+\sigma^2$$

偏差（bias）是模型本身导致的误差，即错误的模型假设所导致的误差，它是模型的预测值的数学期望和真实值之间的差距。
方差（variance）是由于对训练样本集的小波动敏感而导致的误差。它可以理解为模型预测值的变化范围，即模型预测值的波动程度。

公式推导

这里要用到期望和方差的推导

$$y = f + \varepsilon$$

$$\begin{aligned}
\mathrm{E}\left(\left(y-\hat{f}\right)^2\right) &= \mathrm{E}\left(y^2+\hat{f}^2-2y\hat{f}\right) \\
&= \mathrm{E}\left(y^2\right)+\mathrm{E}\left(\hat{f}^2\right)-\mathrm{E}\left(2y\hat{f}\right) \\
&= \mathrm{Var}\left(y\right)+\mathrm{E}^2\left(y\right)+\mathrm{Var}\left(\hat{f}\right)+\mathrm{E}^2\left(\hat{f}\right)-2f\mathrm{E}\left(\hat{f}\right) \\
&= \mathrm{Var}\left(y\right)+\mathrm{Var}\left(\hat{f}\right)+\left(f^2-2f\mathrm{E}\left(\hat{f}\right)+\mathrm{E}^2\left(\hat{f}\right)\right) \\
&= \mathrm{Var}\left(y\right)+\mathrm{Var}\left(\hat{f}\right)+\left(f-\mathrm{E}\left(\hat{f}\right)\right)^2 \\
&= \sigma^2+\mathrm{Var}\left(\hat{f}\right)+\mathrm{Bias}^2\left(\hat{f}\right)
\end{aligned}$$

4.20 正则化

在机器学习中需要抵抗过拟合问题

$$L(\theta) = \frac{1}{2l}\sum_{i=1}^{l}\left(h_\theta\left(\mathrm{x}_i\right)-y_i\right)^2$$

$$L(\theta) = \frac{1}{2l}\sum_{i=1}^{l}\left(f_\theta\left(\mathrm{x}^i\right)-y^i\right)^2+\frac{\lambda}{2}r(\theta)$$

$$L(\theta) = \frac{1}{2l}\sum_{i=1}^{l}\left(h_\theta\left(\mathrm{x}^i\right)-y^i\right)^2+\frac{\lambda}{2}\|\theta\|^2$$

岭回归-线性回归+L2正则化

$$\min_{\mathrm{w}} \sum_{i=1}^{l}\left(\mathrm{w}^{\mathrm{T}}\mathrm{x}_i - y_i\right)^2 \qquad \frac{\partial L}{\partial w_j} = \sum_{i=1}^{l} 2\left(\sum_{k=1}^{n} w_k x_{ik} - y_i\right) x_{ij} = 0$$

$$\left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)\mathrm{w} = \mathrm{X}^{\mathrm{T}}\mathrm{y} \qquad \mathrm{X} = \begin{bmatrix} \mathrm{x}_1^{\mathrm{T}} \\ \dots \\ \mathrm{x}_l^{\mathrm{T}} \end{bmatrix} \qquad \mathrm{w} = \left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{y}$$

$$\min_{\mathrm{w}} \sum_{i=1}^{l}\left(\mathrm{w}^{\mathrm{T}}\mathrm{x}_i - y_i\right)^2 + \lambda \mathrm{w}^{\mathrm{T}}\mathrm{w} \qquad \mathrm{w} = \left(\mathrm{X}^{\mathrm{T}}\mathrm{X} + \lambda \mathrm{I}\right)^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{y}$$

LASSO回归-线性回归+L1正则化

$$\min_{\mathrm{w}} \sum_{i=1}^{l}\left(\mathrm{w}^{\mathrm{T}}\mathrm{x}_i - y_i\right)^2 + \lambda \left\|\mathrm{w}\right\|_1$$

- L1正则化：一范数
- L2正则化：二范数

4.21 岭回归

岭回归=线性回归+L2正则化项

岭回归-线性回归+L2正则化

$$\min_{\mathrm{w}} \sum_{i=1}^{l}\left(\mathrm{w}^{\mathrm{T}}\mathrm{x}_i - y_i\right)^2 \qquad \frac{\partial L}{\partial w_j} = \sum_{i=1}^{l} 2\left(\sum_{k=1}^{n} w_k x_{ik} - y_i\right) x_{ij} = 0$$

$$\left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)\mathrm{w} = \mathrm{X}^{\mathrm{T}}\mathrm{y} \qquad \mathrm{X} = \begin{bmatrix} \mathrm{x}_1^{\mathrm{T}} \\ \dots \\ \mathrm{x}_l^{\mathrm{T}} \end{bmatrix} \qquad \mathrm{w} = \left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{y}$$

$$\min_{\mathrm{w}} \sum_{i=1}^{l}\left(\mathrm{w}^{\mathrm{T}}\mathrm{x}_i - y_i\right)^2 + \lambda \mathrm{w}^{\mathrm{T}}\mathrm{w} \qquad \mathrm{w} = \left(\mathrm{X}^{\mathrm{T}}\mathrm{X} + \lambda \mathrm{I}\right)^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{y}$$

LASSO回归-线性回归+L1正则化

$$\min_{\mathrm{w}} \sum_{i=1}^{l}\left(\mathrm{w}^{\mathrm{T}}\mathrm{x}_i - y_i\right)^2 + \lambda \left\|\mathrm{w}\right\|_1$$

如果一个矩阵严格对角占优，那么它可逆。

如果一个矩阵严格对角占优，那么它是可逆的。

LASSO回归=线性回归+L1正则化项

LASSO回归-线性回归+L1正则化

$$\min_{\mathrm{w}} \sum_{i=1}^{l}\left(\mathrm{w}^{\mathrm{T}}\mathrm{x}_i - y_i\right)^2 + \lambda \left\|\mathrm{w}\right\|_1$$

4.22 本集总结

这节课的内容事比较多的，需要好好消化。

1. 分类
    1. 有监督
        1. 分类
            1. 全
            2. 判别分析
        2. 回归
    2. 无监督
    3. 强化

2. 回归和误差
    1. 精度和赵化率
    2. ROC
    3. CV
3. 欠拟合过拟合
    1. 变差和方差
    2. #