四川大學
SICHUAN UNIVERSITY

机器学习-第五章 机器学习实践

教师：胡俊杰 副教授

邮箱：hujunjie@scu.edu.cn

1.贝叶斯定理

$$P(X,Y) = P(Y|X)P(X)$$

$$P(X,Y) = P(X|Y)P(Y)$$

$$\longrightarrow P(Y|X)P(X) = P(X|Y)P(Y) \longrightarrow P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} = \frac{P(X|Y)P(Y)}{P(X|Y)P(Y) + P(X|\bar{Y})P(\bar{Y})}$$

全概率公式

1.贝叶斯方法-背景知识

贝叶斯分类：

贝叶斯分类是一类分类算法的总称，这类算法均以贝叶斯定理为基础，故统称为贝叶斯分类

先验概率
Prior probability:

根据以往经验和分析得到的概率，记为$P(Y)$。是在观测数据前，表达事件不确定性的概率分布，其代表经验知识，与观测数据无关

后验概率
Posterior probability:

给定观测数据X后，对事件Y发生概率的更新，记为$P(Y|X)$，它反映了在获取新数据X后，调整对Y发生可能性的评估

1.贝叶斯方法-一个简单示例

- 假设某种疾病在所有人群中的感染率是0.1%
- 医院现有的技术对于该疾病检测准确率为 99%（已知患病情况下， 99% 的可能性可以检查出阳性；正常人 99% 的可能性检查为正常。）

问：从人群中随机抽一个人去检测，医院给出的检测结果为阳性，那么这个人实际得病的概率是多少?

99%？

1.贝叶斯方法-一个简单示例

Y: 某人患有该疾病

X: 医院检测结果为阳性（检测结果显示患病）

- 医院现有的技术对于该疾病检测准确率为 99%：$P(X|Y) = 99\%$

问：从人群中随机抽一个人去检测，医院给出的检测结果为阳性，那么这个人实际得病的概率是多少?

即求$P(Y|X)$

贝叶斯公式

1.贝叶斯方法-一个简单示例

X:医院检测结果为阳性（检测结果显示患病）　　Y:某人患有该疾病

贝叶斯公式：　$P(Y|X)=\frac{P(X|Y)P(Y)}{P(X)}=\frac{P(X|Y)P(Y)}{P(X|Y)P(Y)+P(X|\bar{Y})P(\bar{Y})}$

- 假设某种疾病在所有人群中的感染率是0.1%: $P(Y)=0.1\%, P(\bar{Y})=99.9\%$
- 医院现有的技术对于该疾病检测准确率为 99%: $P(X|Y)=99\%, P(X|\bar{Y})=\frac{P(X,\bar{Y})}{P(\bar{Y})}=\frac{0.01}{0.999}\approx 1\%$

错检/误诊

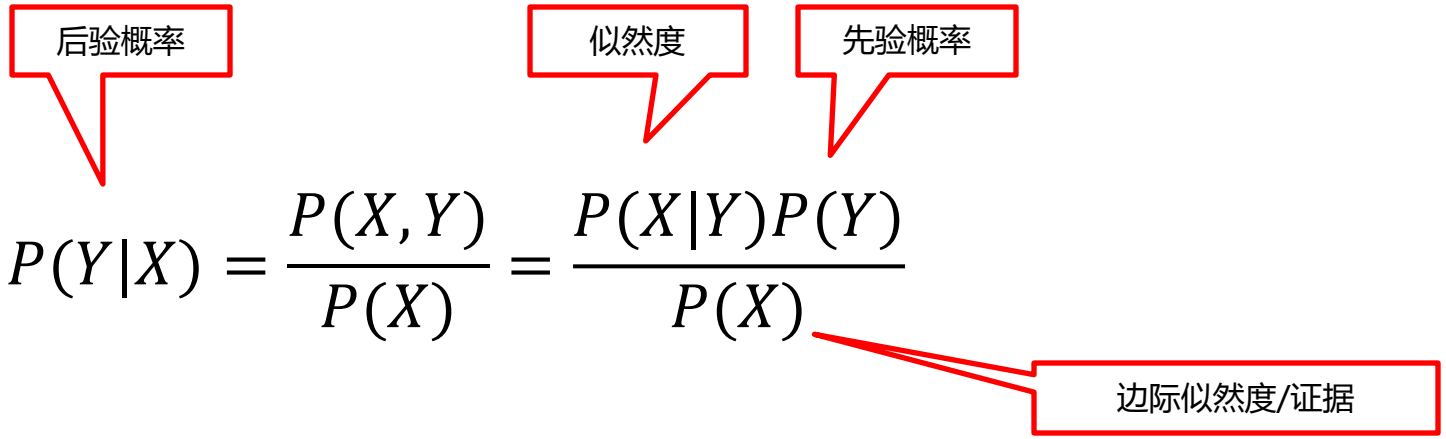
$$P(Y|X)=\frac{0.99*0.001}{0.99*0.001+0.01*0.999}\approx 0.09$$

从人群中随机抽一个人去检测，医院给出的检测结果为阳性，实际真实得病的概率为9%

联系生活中的贝叶斯

1. 贝叶斯方法

贝叶斯公式



朴素贝叶斯法是典型的生成学习方法。生成方法由训练数据学习联合概率分布 $P(X,Y)$，然后求得后验概率分布$P(Y|X)$。

具体来说，利用训练数据学习$P(X|Y)$和$P(Y)$的估计，得到联合概率分布：

$P(X,Y) = P(X|Y)P(Y)$

2.朴素贝叶斯原理

判别模型和生成模型

监督学习模型可分为
判别模型（Discriminative model）和**生成模型**（Generative model）

判别模型（Discriminative model）	生成模型（Generative model）
由数据直接学习决策函数$Y = f(X)$或者条件概率分布$P(Y\|X)$ 的模型，即判别模型。基本思想是在有限样本条件下建立判别函数，不考虑样本的产生模型，直接研究预测模型。 即：直接估计$P(Y\|X)$	由训练数据学习联合概率分布 $P(X,Y)$，然后求得后验概率分布$P(Y\|X)$。具体来说，利用训练数据学习$P(X\|Y)$和$P(Y)$的估计，得到联合概率分布：$P(X,Y) = P(Y)P(X\|Y)$，再利用它进行分类。 即：估计$P(X\|Y)$ 然后推导$P(Y\|X)$
线性回归、逻辑回归、感知机、决策树、支持向量机……	朴素贝叶斯、HMM……

2.朴素贝叶斯原理

$$P(Y=c_k|X=x)=\frac{P(X=x|Y=c_k)P(Y=c_k)}{P(X=x)}$$

$$P(X=x|Y=c_k)=P\left(X^{(1)}=x^{(1)},\dots,X^{(n)}=x^{(n)}\middle|Y=c_k\right),k=1,2,\dots,K$$

- 假设$x^{(j)}$可能的取值有S_j个，$j=1,2,\dots n$，Y可能值有K个，则$P(X=x|Y=c_k)$的可能情况有 $K\prod_{j=1}^{n}S_j$种，复杂度高
- 若假设在类别确定的条件下，各特征相互独立，则

$$P(X=x|Y=c_k)=P\left(X^{(1)}=x^{(1)},\dots,X^{(n)}=x^{(n)}\middle|Y=c_k\right)$$

简化问题

$$=\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}|Y=c_k)$$

2.朴素贝叶斯原理

$$P(X=x|Y=c_k)=P\left(X^{(1)}=x^{(1)},\dots,X^{(n)}=x^{(n)}\middle|Y=c_k\right)=\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}|Y=c_k)$$

贝叶斯公式：$$P(Y=c_k|X=x)=\frac{P(X=x|Y=c_k)P(Y=c_k)}{P(X=x)}=\frac{P(X=x|Y=c_k)P(Y=c_k)}{\sum_i P(X=x|Y=c_i)P(Y=c_i)}$$

$$P(Y=c_k|X=x)=\frac{P(Y=c_k)\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}|Y=c_k)}{\sum_i P(Y=c_i)\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}|Y=c_i)}$$

2.朴素贝叶斯原理

$$y = \underset{c_k}{\operatorname{argmax}} \frac{P(Y = c_k) \prod_{j=1}^{n} P(X^{(j)} = x^{(j)} | Y = c_k)}{\sum_i P(Y = c_i) \prod_{j=1}^{n} P(X^{(j)} = x^{(j)} | Y = c_i)}$$

- 对任意的c_k而言，以上公式的分母均相等，因此

$$y = \underset{c_k}{\operatorname{argmax}} P(Y = c_k) \prod_{j=1}^{n} P(X^{(j)} = x^{(j)} | Y = c_k)$$

2.朴素贝叶斯原理

$$y = \underset{c_k}{\operatorname{argmax}} P(Y = c_k) \prod_{j=1}^{n} P(X^{(j)} = x^{(j)} | Y = c_k)$$

c_k类样本的数目

$$P(Y = c_k) = \frac{\sum_{i=1}^{N} I(y_i = c_k)}{N}, k = 1,2, \dots, K$$

N: 训练样本数目

$$I(y_i = c_k) = \begin{cases} 1, if\ y_i = c_k \\ 0, else \end{cases}$$

Indicator function (指示函数)

属于c_k类，且输入特征为a_{jl}样本的数目

$$P\left(X^{(j)} = a_{jl} \middle| Y = c_k\right) = \frac{\sum_{i=1}^{N} I(x_i^{(j)} = a_{jl}, y_i = c_k)}{\sum_{i=1}^{N} I(y_i = c_k)}$$

c_k类样本的数目

$j = 1,2, \dots, n$ 每个样本共有n维特征

$l = 1,2, \dots, S_j$ 第j维特征可能有S_j种取值

$k = 1,2, \dots, K$ 共有K个类别

$x_i^{(j)}$:第i个样本的第j个特征　　a_{jl}: 第j个特征可能取的第l个值

1.数据集划分

01 数据集划分

02 评价指标

03 正则化

1.数据集划分

- 数据集$T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$，共包含$N$个样本，其中$x_i$代表第$i$个输入的样本，$y_i$代表与$x_i$对应的标签

例1：根据患者的影像数据来判断肿瘤的良性或恶性？

x_i：第i个患者的影像

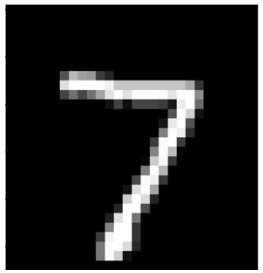
y_i：良性/恶性

x_i

例2：根据手写体数字图像，识别图像中的数字

x_i：第i张手写体数字的图像

y_i：0~9



x_i

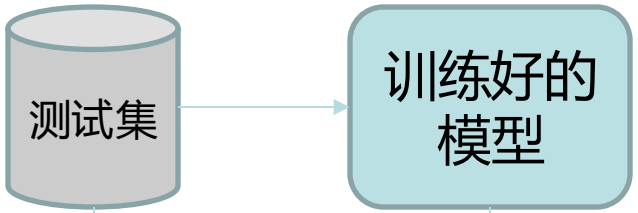
使用部分数据训练模型，使用部分数据评价模型的效果

1.数据集划分

训练集（Training Set）：训练模型所使用到的数据，通过该部分数据确定模型所包含的各学习参数

测试集（Test Set）： 测试已经训练好的模型的性能

训练阶段
(如训练手写体数字识别模型)



实用阶段
(将训练好的手写体数字模型用于真实应用场景)

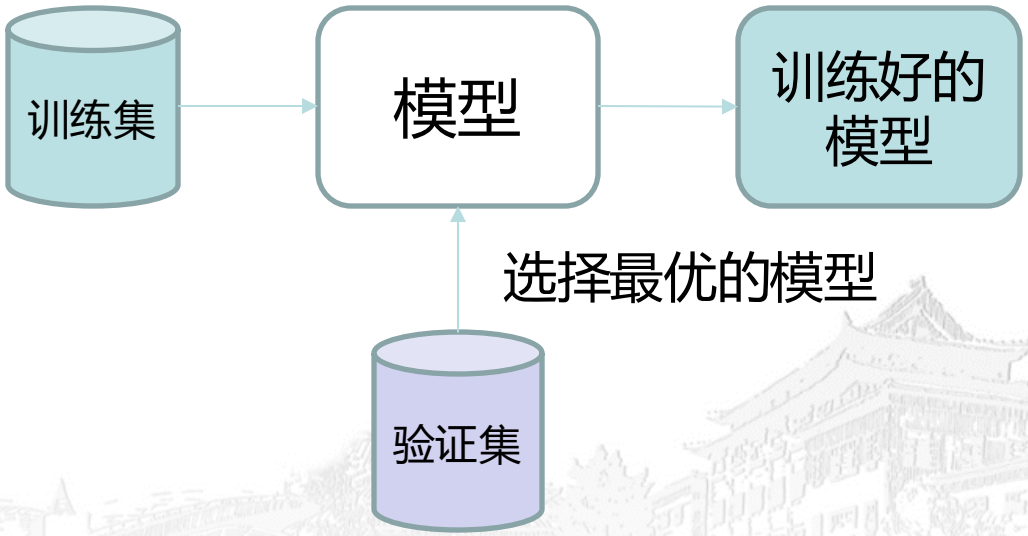
训练过程中如何挑选模型?

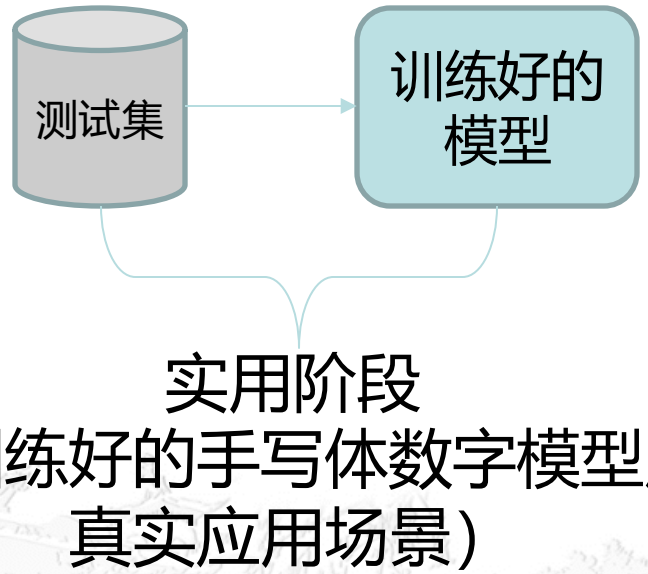
1.数据集划分

训练集（Training Set）：训练模型所使用到的数据，通过该部分数据确定模型所包含的各学习参数

验证集（Validation Set）：有时也叫做开发集（Dev Set），用来做模型选择（model selection），评价模型的训练效果

测试集（Test Set）： 测试已经训练好的模型的性能





1.数据集划分

训练集（Training Set）：训练模型所使用到的数据，通过该部分数据确定模型所包含的各学习参数

验证集（Validation Set）：有时也叫做开发集（Dev Set），用来做模型选择（model selection），评价模型的训练效果

测试集（Test Set）： 测试已经训练好的模型的性能

- 三者划分：训练集（80%）、验证集（10%）、测试集（10%）
- 实际应用中，训练集/验证集/测试集的具体比例可调整
- 如果只划分训练集和验证集，通常训练集（80%），验证集（20%）

1.数据集划分

第1步. 准备训练数据集 $D = \{(x, y)\}$

第2步. 随机初始化$(w_0, w_1, w_2, \ldots, w_n)$， 设置学习率 α.

第3步. 从 D中选择b个训练样本， $(x_i, y_i) \in D^b$

$$\frac{\partial J(w)}{\partial w_j} \leftarrow \frac{\partial J(w)}{\partial w_j} + (h(x_i) - y_i)x_i^{(j)}$$ //计算并累积各样本的梯度

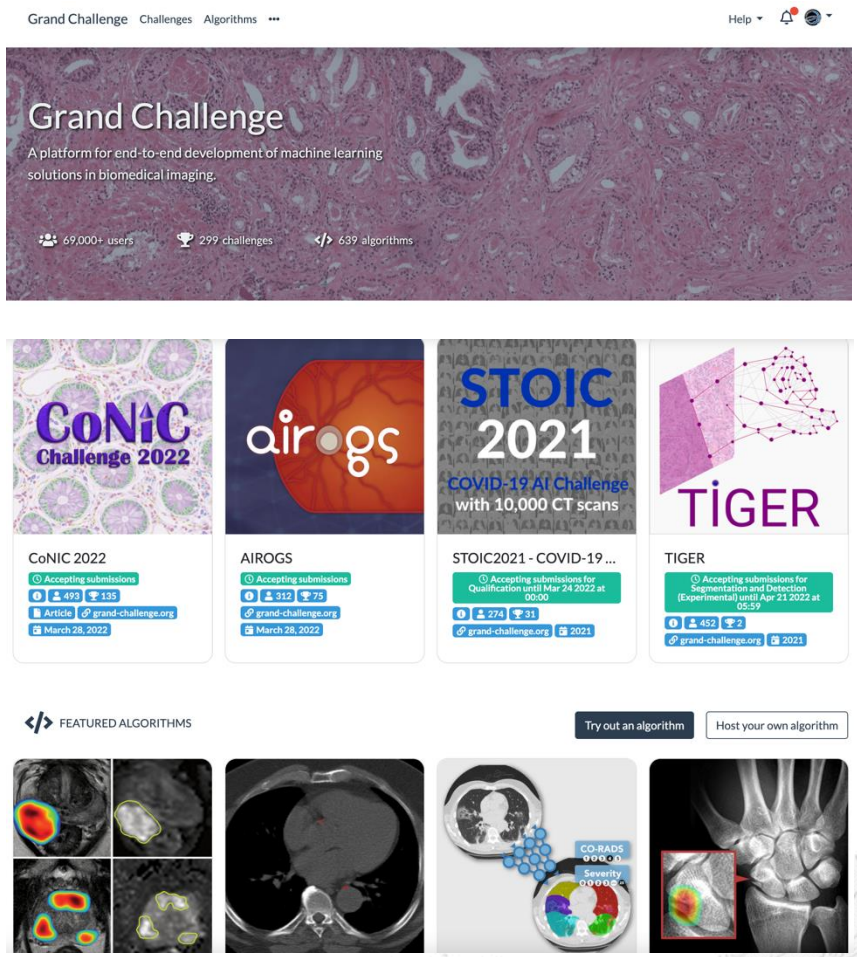
第4步. 更新参数

$$w_j := w_j - \alpha \frac{1}{b} \frac{\partial J(w)}{\partial w_j}$$

第5步. 继续第3步，直到模型收敛.

- 使用验证集判断模型是否已收敛

1.数据集划分



- 只给出训练集（数据+标签），测试集不公开
  - 以docker或其他方式提交模型，由主办方验证测试集性能指标
- 给出训练集（数据+标签），以及测试集（仅数据）
  - 提交模型对于测试集的预测结果，由主办方计算测试集性能指标

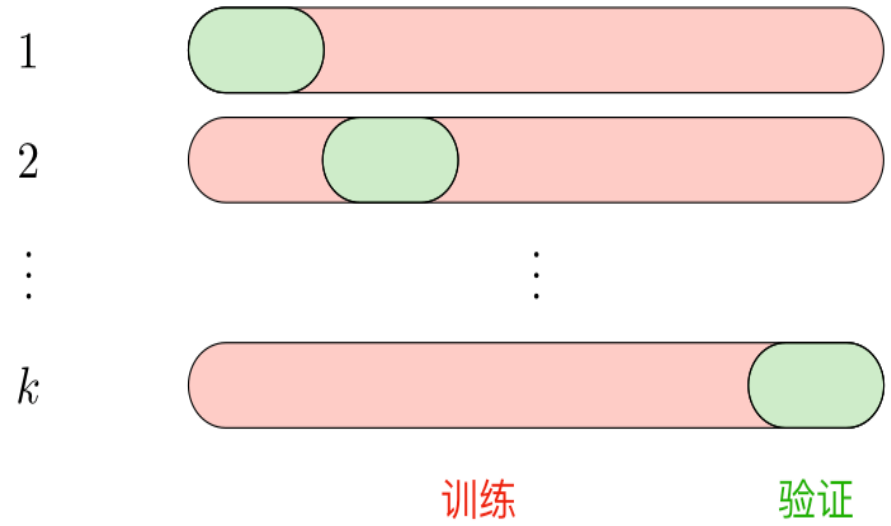
Test B榜 | Test A榜

#	团队名称	成员	最好成绩	检测结果	分割结果
1	Machine Intelligence Lab		0.8706	0.989	0.7521
2	965728310		0.8684	0.9903	0.7466
3	Looking		0.8665	0.9872	0.7458
4	DeepSeg		0.865	0.9859	0.7441
5	Menelvagor		0.8646	0.9895	0.7398

交叉验证

k交折（k-fold）

子集　　数据集划分　　验证集误差　　平均验证集误差



ϵ_1

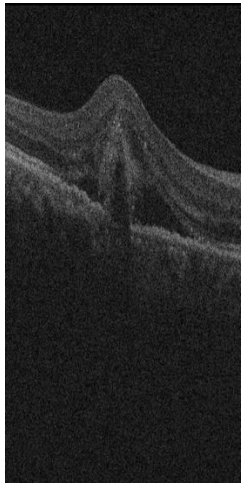
ϵ_2

ϵ_k

$$\frac{\epsilon_1 + \ldots + \epsilon_k}{k}$$

- 划分不同的训练集/验证集，训练得到k个模型，每个模型对应的误差为ϵ
  - 此处的误差可以是验证集的代价函数值，或验证集的其他性能指标（如分类准确率）
- 用k个模型分别对交叉验证集计算得出平均交叉验证误差
- 选择验证集误差最小的模型用于真实应用场景

不平衡数据的处理

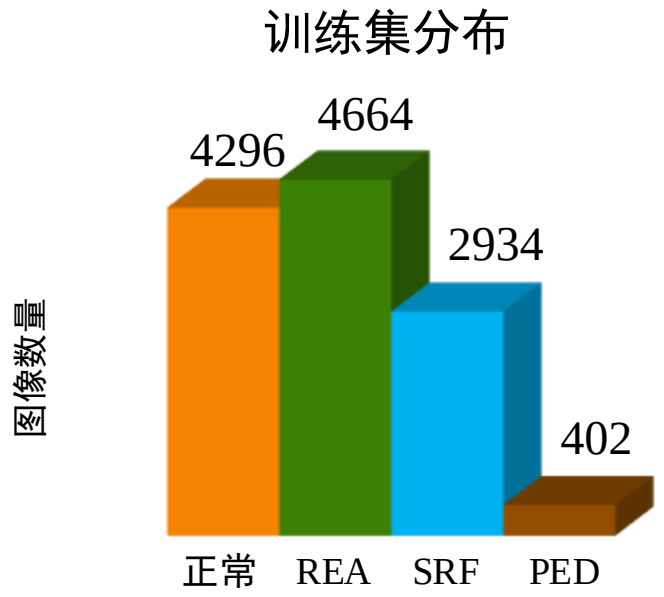


输入图像（OCT）

任务I：黄斑水肿病变分类 → 是否正常、REA、SRF、PED？

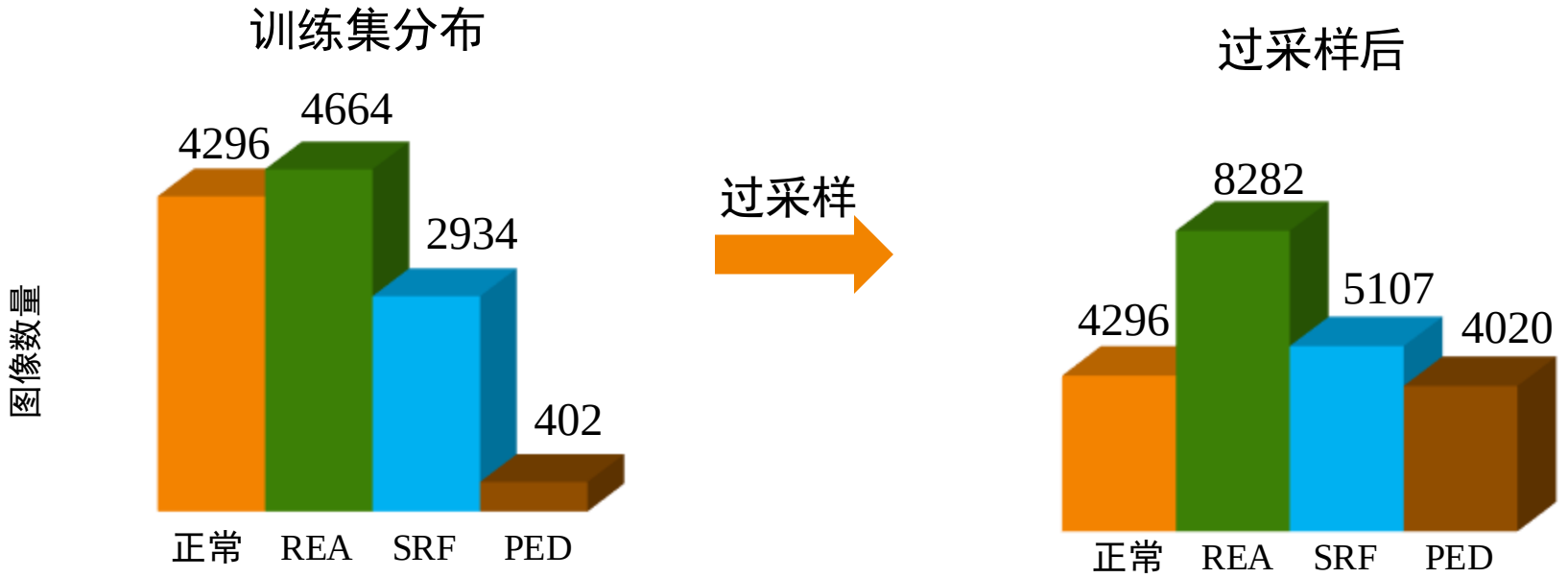
任务II：黄斑水肿病变分割 → 

识别各病灶区域



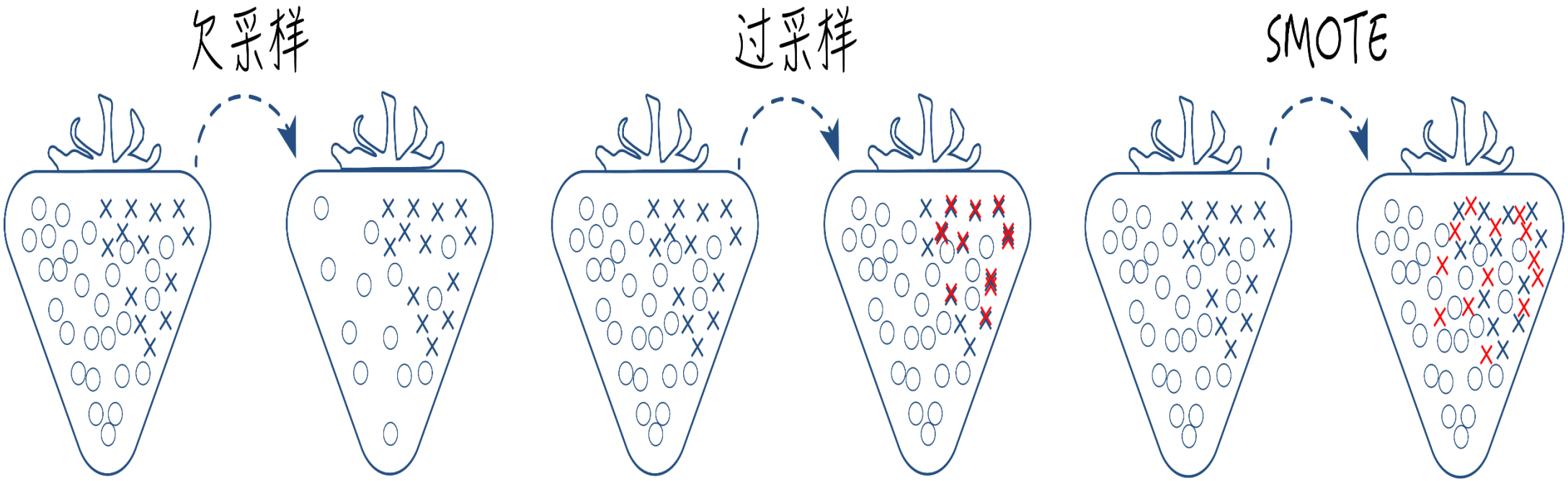
- 数据不平衡是指数据集中各类样本数量不均衡的情况
- 不加处理的话，模型会倾向于预测正常和REA类别，忽略PED类别

不平衡数据的处理



- 对数量较少的类别采用有放回的方式重复采样。由于该任务是一个多标签任务（一张图像对应多个标签），对SRF、PED类别过采样也将增加REA类别的数量
- 实际应用中，应尽可能保证各类别样本数量相近

不平衡数据的处理



- 过采样方法仅是对数量较少类别样本的简单复制，如何增加样本的多样性
- SMOTE: Synthetic Minority Over-sampling TEchnique

[*] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of artificial intelligence research, 2002, 16: 321-357.

2.评价指标

01 数据集划分

02 评价指标

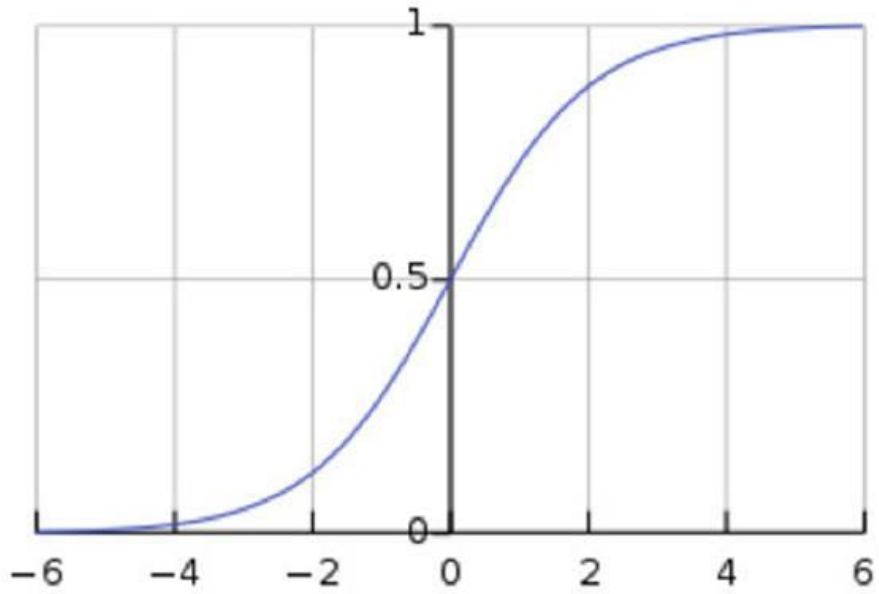
03 正则化

评价指标

Logistic回归

$$y = \sigma(z) = \frac{1}{1+e^{-z}} \qquad z=w^T x + b$$

- w和b是可学习参数
- x是输入数据， y是模型输出



假设$y = 0.6$

- 当判定的阈值设置为0.5时，模型预测为正类别
- 当判断的阈值设置为0.7时，模型预测为负类别

模型的预测类别与阈值的设置有关，不同的阈值将导致不同的分类结果

评价指标

Positive: 阳性

Negative: 阴性

- 预测类别（2类）：阳性/阴性
- 真实类别（2类）：阳性/阴性

如何准确描述模型预测正确与否?

评价指标

模型是否预测正确
- 预测正确（True）
- 预测错误（False）

模型预测类别
- 阳性
- 阴性

1. **真阳（True Positive,TP）**： 预测正确，预测为阳性，真实为阳性
2. **真阴（True Negative,TN）**：预测正确，预测为阴性，真实为阴性
3. **假阳（False Positive,FP）**： 预测错误，预测为阳性，真实为阴性
4. **假阴（False Negative,FN）**：预测错误，预测为阴性，真实为阳性

评价指标

1. **真阳（True Positive,TP）：** 预测为阳性，真实为阳性
2. **真阴（True Negative,TN）：** 预测为阴性，真实为阴性
3. **假阳（False Positive,FP）：** 预测为阳性，真实为阴性
4. **假阴（False Negative,FN）：** 预测为阴性，真实为阳性

混淆矩阵（confusion matrix）

		预测类别（预测）	
		Positive	Negative
真实类别（标签）	Positive	TP	FN
	Negative	FP	TN

- 横轴为真实类别
- 纵轴为预测类别

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

评价指标

- 假设有100人，其中1人为阳性患者，99人健康
- 现有预测该病种的模型，该模型只输出阴性

请给出模型的混淆矩阵

评价指标

		预测类别（预测）	
		Positive	Negative
真实类别（标签）	Positive	0	**1**
	Negative	**0**	**99**

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{99}{100}$$

评价指标

		预测类别（预测）	
		Positive	Negative
真实类别（标签）	Positive	0	**1**
	Negative	**0**	**99**

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{99}{100}$$

- 准确率（Accuracy）指标可客观地度量类别平衡时模型的性能
- 当类别不平衡时，需要引入新的度量指标来评价模型的效果

评价指标

混淆矩阵（**confusion matrix**）

		预测	
		Positive	Negative
标签	Positive	TP	FN
	Negative	FP	TN

召回率：$Recall = \frac{TP}{TP+FN}$

- 针对真实的阳性样本，即在全体阳性样本中，模型预测出的比例

精准率：$Precision = \frac{TP}{TP+FP}$

- 针对模型预测为阳性的样本，即在全体模型预测为阳性的样本，真实阳性样本所占的比例

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

- 同时结合了召回率和精准率特点的综合性评价指标

评价指标

混淆矩阵（confusion matrix）

		预测	
		Positive	Negative
标签	Positive	TP	FN
	Negative	FP	TN

注意：召回率、精准率、$F1$指标均只针对二分类。若是多分类，则采用**One-vs-Rest**的策略，将某一类视为阳性（positive），其余类别视为阴性（negtive）

召回率：$Recall = \frac{TP}{TP+FN}$

- 针对真实的阳性样本，即在全体阳性样本中，模型预测出的比例

精准率：$Precision = \frac{TP}{TP+FP}$

- 针对模型预测为阳性的样本，即在全体模型预测为阳性的样本，真实阳性样本所占的比例

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

- 同时结合了召回率和精准率特点的综合性评价指标

评价指标

有100张照片，其中，猫的照片有60张，狗的照片是40张。

输入这100张照片进行二分类识别，找出这100张照片中的所有的猫。

识别结果的混淆矩阵

		预测	
		Positive	Negative
标签	Positive	TP=40	FN=20
	Negative	FP=10	TN=30

请给出Accuracy、Recall和Precision

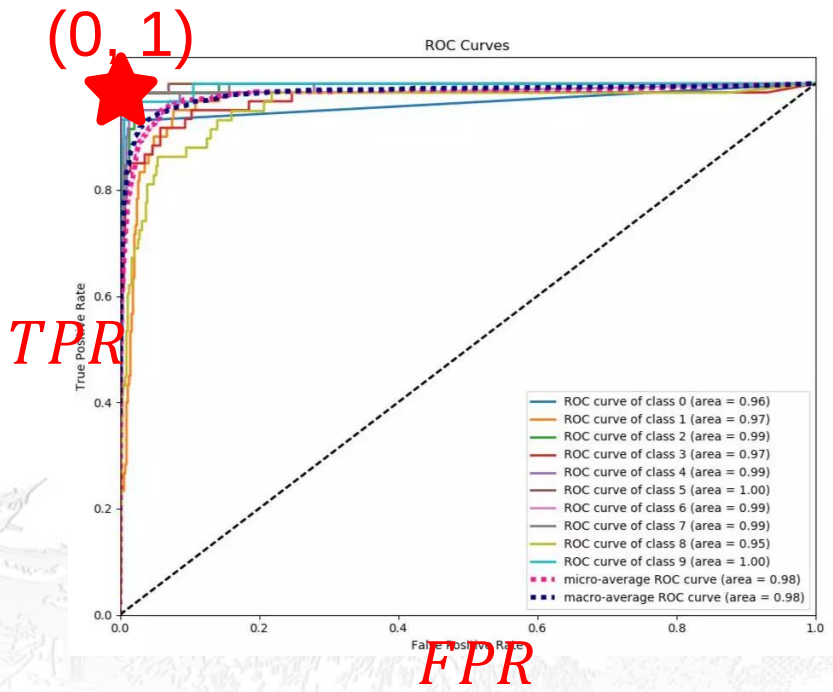
评价指标

ROC曲线， Receiver Operating Characteristic curve，也称为受试者工作特征曲线

$$FPR = \frac{FP}{FP + TN}$$ 假阳率：针对所有阴性样本，被错误预测为阳性的比例 ↓ 越小越好

$$TPR = \frac{TP}{TP + FN}$$ 真阳率：针对所有阳性样本，被正确预测为阳性的比例 ↑ 越大越好

- 随着阈值的不同，FPR和TPR都在同步变化，(FPR, TPR)所构成的曲线则称为ROC曲线
- ROC曲线与坐标轴围成的面积，称为AUC（Area Under Curve, 曲线下面积），面积越大，则模型性能越好

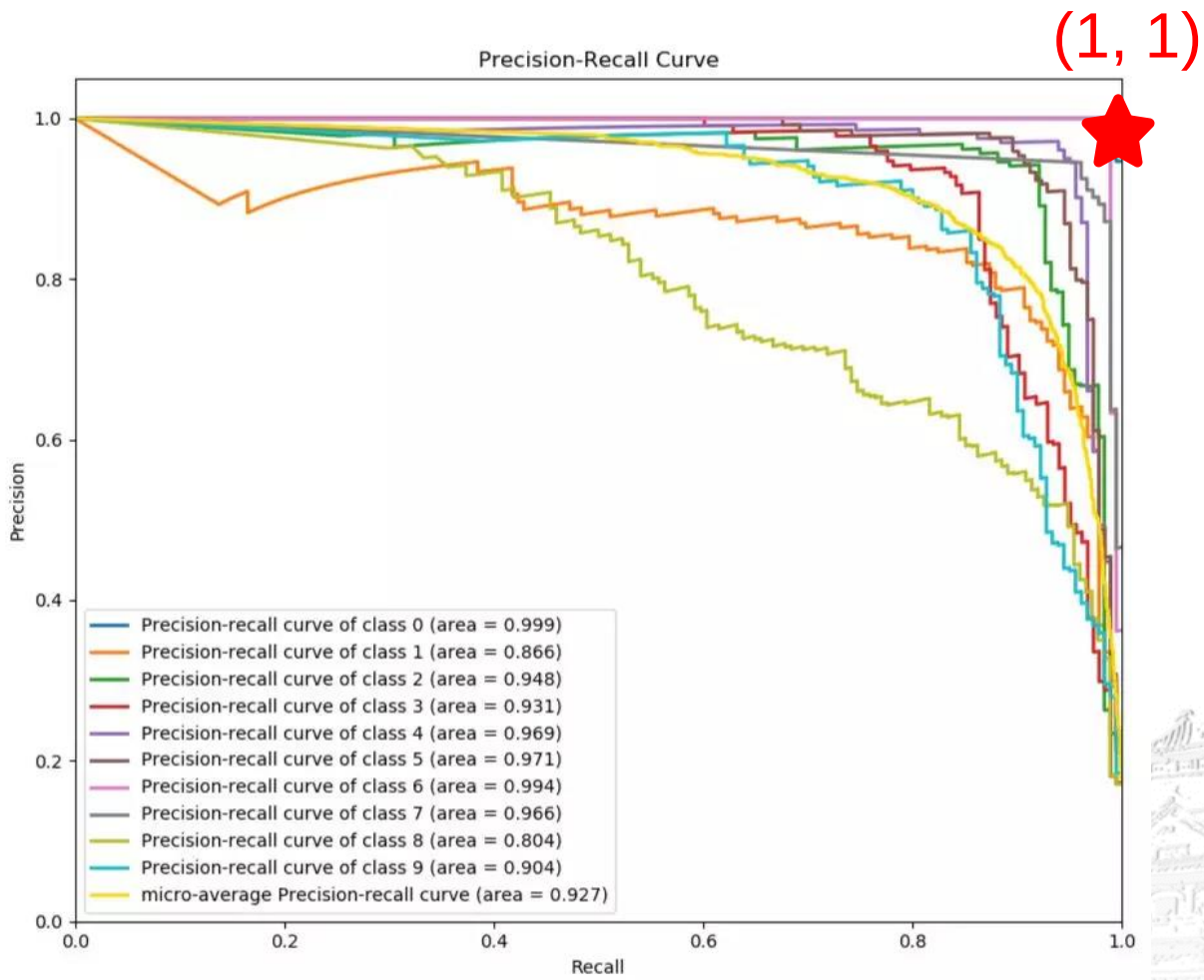


评价指标

PR（Precision-Recall）曲线

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$



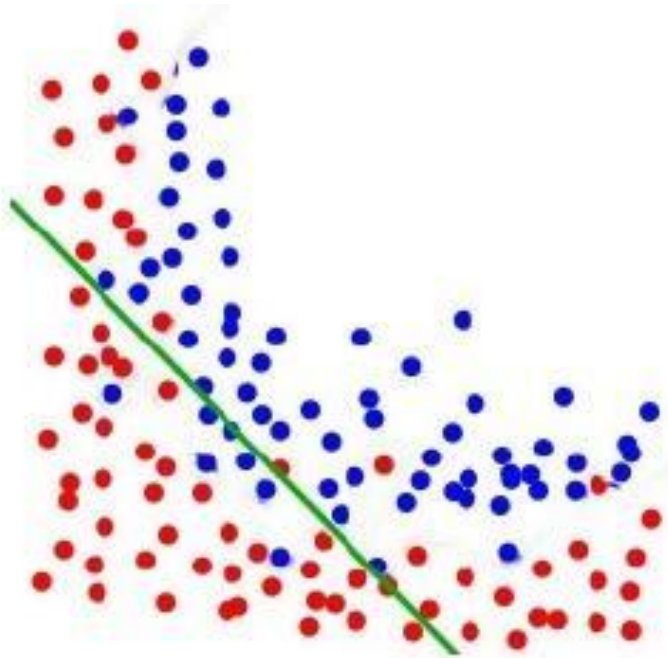
3.正则化、偏差和方差

01 数据集划分

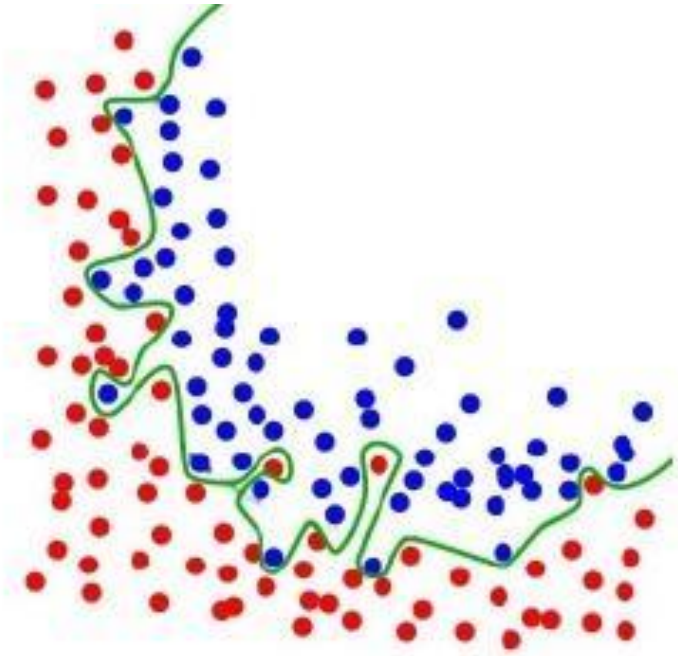
02 评价指标

03 正则化

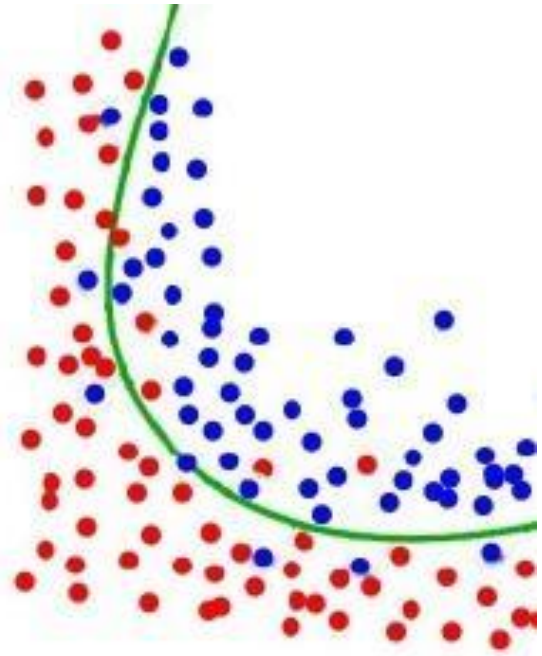
过拟合和欠拟合



欠拟合



过拟合



正合适

正则化（Regularization）

正则化系数

L_1正则化：$J(w)=\frac{1}{2}\sum_{i=1}^{m}(h(x_i)-y_i)^2+\lambda\sum_{j=1}^{n}|w_j|$，Lasso Regression（Lasso回归）

L_2正则化：$J(w)=\frac{1}{2}\sum_{i=1}^{m}(h(x_i)-y_i)^2+\lambda\sum_{j=1}^{n}w_j^2$，Ridge Regression （岭回归）

Elastic Net：$J(w)=\frac{1}{2}\sum_{i=1}^{m}(h(x_i)-y_i)^2+\lambda(\rho\cdot\sum_{j=1}^{n}|w_j|+(1-\rho)\cdot\sum_{j=1}^{n}w_j^2)$

(弹性网络)

比例系数

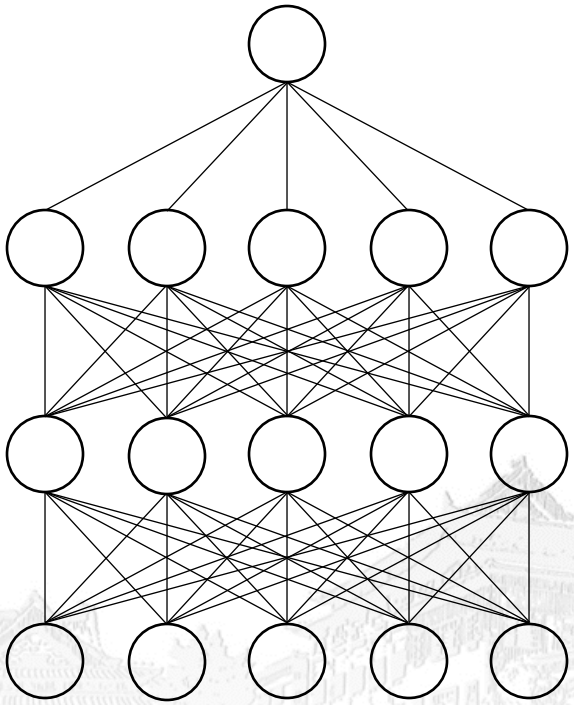
其中：

- λ为正则化系数，调整正则化项与训练误差的比例，λ>0。
- 1≥ρ≥0为比例系数，调整L_1正则化与L_2正则化的比例。

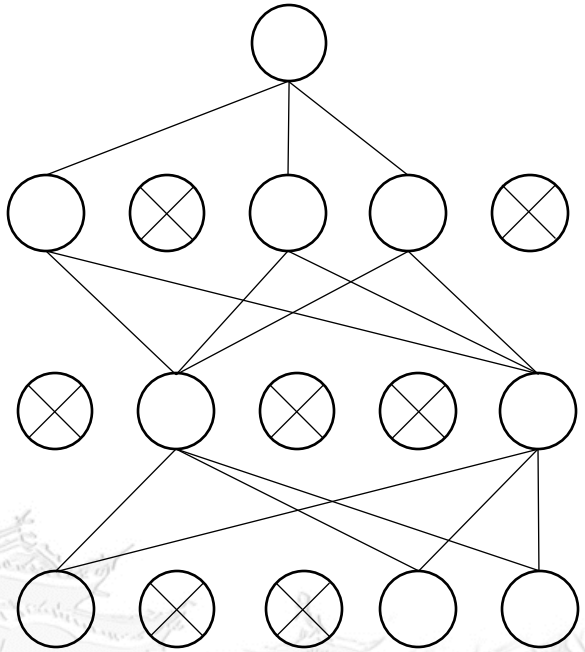
正则化

Dropout（深度神经网络训练过程中常用的正则化）

- 训练过程：以概率p随机地禁止/激活每个神经元（伯努利分布）
- 测试过程：保留全体神经元，但激活值强度乘以p



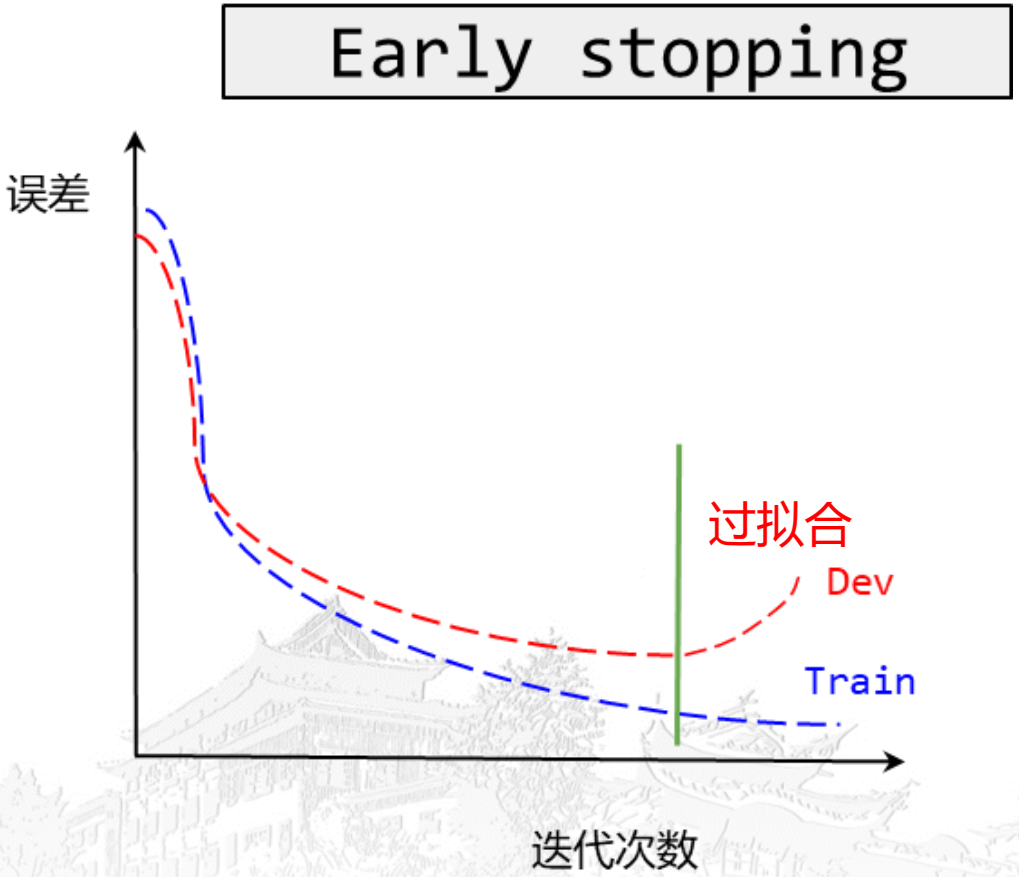
标准的深度神经网络



作用了Dropout之后的深度神经网络

正则化

Early stopping代表提早停止训练模型

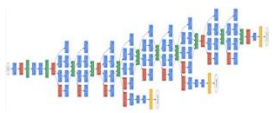


正则化

- 数据收集困难
- 标注人力成本高

大数据

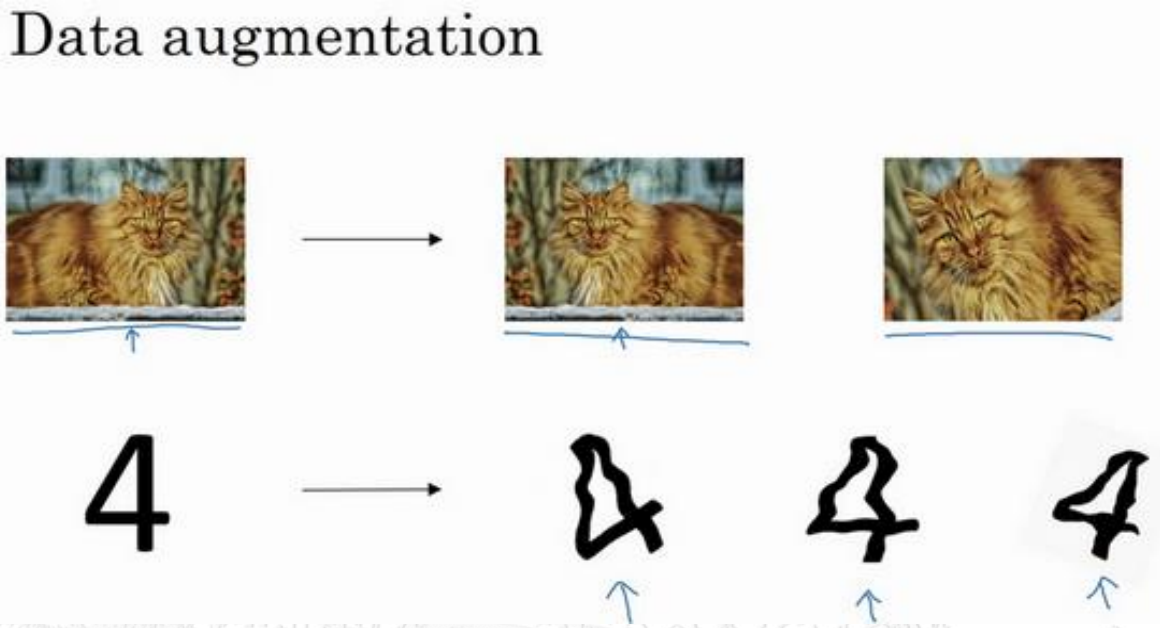
大模型



大算力

数据增广：人为增强数据的多样性

- 颜色空间：亮度、灰度、对比度等
- 几何空间：旋转、平移、缩放、弹性形变等



作业

		预测			
		类1	类2	类3	类4
标签	类1	a_{11}	a_{12}	a_{13}	a_{14}
	类2	a_{21}	a_{22}	a_{23}	a_{24}
	类3	a_{31}	a_{32}	a_{33}	a_{34}
	类4	a_{41}	a_{42}	a_{43}	a_{44}

- 请写出类1、类2、类3、类4的Recall和Precision
- 请解释为什么要引入Recall和Precision
- 请解释Recall值高/低的含义，以及Precision值高/低的含义