第二章 模型评估与选择

误差与过拟合

- `error rate` 错误率 = 分类错误的样本数占总样本数的比例：$E = a/m$
- `accuracy` 精度 = 1 - 错误率：$1 - a/m$
- 误差：

误差名	使用范围
误差	实际输出与样本真实输出之间
训练误差/经验误差	在训练集上的误差
泛化误差	在新样本上的误差

- 过拟合 `overfitting`
- 欠拟合 `underfitting`

  比如对一张带有锯齿的绿色树叶进行学习，过拟合的结果是认为所有树叶都必须带有锯齿，欠拟合的结果是绿的的都为树叶。

评估方法

对机器学习的泛化误差进行评估，需要一个 **测试集** ，用测试集上的 **测试误差** 多位泛化误差的近似。**测试样本尽可能不出现在训练集中**

留出法 hold-out

将数据集 D 划分为两个互斥集合训练集 S 和测试集 T, $D = S \cup T, S \cap T = \varnothing$

- 常规划分比例：2/3 ～ 4/5
- 两个集合尽可能保持数据分布一致
- 由于划分的随机性，单次的留出法结果往往不够稳定，需多次随机划分，重复实验取平均值

交叉验证 cross validation

将数据集划分为 k 个大小相似的互斥子集 $D = D_1 \cup D_2 \cup \ldots \cup D_k, D_i \cap D_j = \varnothing (i \neq j)$

- 子集数据分布保持一致
- 进行k次训练和测试，自身作为测试集，k-1 个座位训练集
- 得到 k 次结果取平均
- 需多次随机划分，重复 p 次实验取平均值，叫 **p次k折交叉验证**
- 当 **k = 样本数** 时为 **留一法LOO**，Leave-One-Out

自助法 bootstrapping

以上两种方法评估模型使用的训练集比 D 小，于是训练样本不同会导致估计偏差，自助法为了减小这一影响

- D 有 m 个样本，从中抽取一个拷贝到 D' 并放回 D (下次采样还有可能被采样到)，在 m 次采样中样本始终不被采样的概率极限

$$\lim_{m\mapsto\infty}(1-\frac{1}{m})^m \mapsto \frac{1}{e} \approx 0.368$$

- 所以 D 中大约有36.8%的样本没有出现在 D' 中，将 D' 作为训练集，$D-D'$ 作为测试集
- **包外估计** 有 3/1 的没有在训练集中出现的样本用于测试

自助法在数据集较小，难以有效划分训练集/测试集时很有用，数据集足够多时前两者比较好。

调参

`parameter tuning`，学习算法中有许多参数 `parameter` 需要设定。

性能评估

有了评估方法，我们还需要一个衡量模型泛化能力的评估标准 `performance measure`。

均方误差 MSE

回归中常用的性能度量方法

- 样本集 $D=\{(x_1,y_1),(x_2,y_2),\dots,(x_m,y_m)\}$
- 预测结果 $f(x)$ 与真实标记 y

均方误差

$$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}(f(x_i)-y_i)^2$$

错误率与精度

错误率定义

$$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}(f(x_i)\neq y_i)$$

精度

$$\begin{aligned}acc(f;D)&=\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}(f(x_i)=y_i)\\&=1-E(f;D)\end{aligned}$$

查准率、查全率、F1

查准率 `precision`: 检索出的信息中有多少比例用户感兴趣；

查全率 `recall`: 用户感兴趣的信息中有多少被检索出来；

对于二分类问题：**样例总数 = TP + FP + TN + FN**

	实际值	预测值	全称
TP	Positive	Positive	True Positive
FP	Negative	Positive	False Positive
FN	Positive	Negative	False Negative
TN	Negative	Negative	True Negative

分类结果 **混淆矩阵** 如下表所示

真实情况	预测结果	
	正例	反例
正例	TP（真正例）	FN（假反例）
反例	FP（假正例）	TN（真反例）

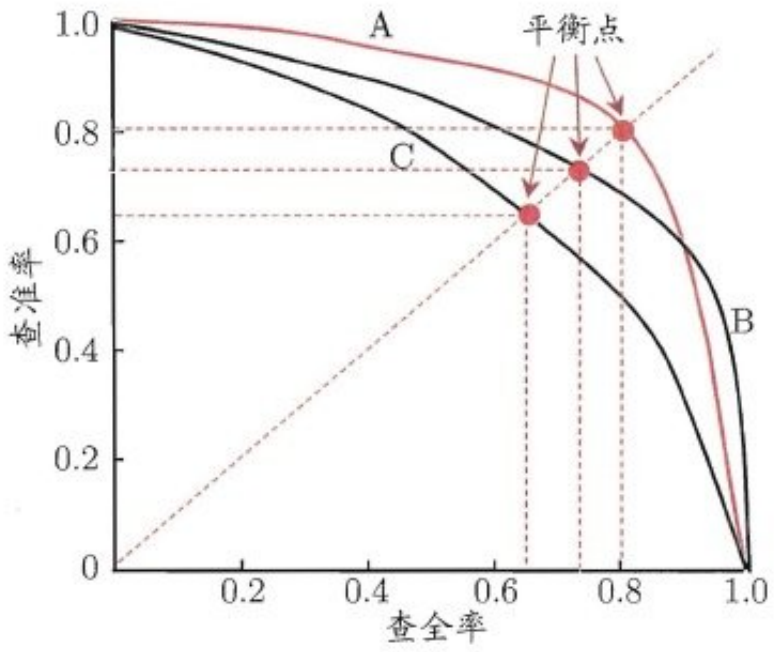
查准率

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

查全率

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

可以看出两者是矛盾量，一般来说P高往往R低。两者关系用 **P-R曲线** 描述，平衡点 **BEP**



F1度量

$$F1=\frac{2\times P\times R}{P+R}=\frac{2\times TP}{m_{(样例总数)}+TP-TN}$$

对于查准率与查全率的重视程度不同，可以设定偏好 β，$\beta=1$ 为标准F1；$\beta>1$ 查全率影响更大；$0<\beta<1$ 查准率影响更大。

$$F_\beta=\frac{(1+\beta^2)\times P\times R}{(\beta^2\times P)+R}$$

多二维混淆矩阵

对于 n 个二分类混淆矩阵综合考察查全率和查准率的方法：

1. 计算个混淆矩阵的 P,R，再求平均，得到宏查准率、宏查全率、宏F1：

$$macro-P=\frac{1}{n}\sum_{i=1}^{n}P_i$$
$$macro-R=\frac{1}{n}\sum_{i=1}^{n}R_i$$
$$macro-F1=\frac{2\times macro-P\times macro-R}{macro-P+macro-R}$$

2. 对混淆矩阵各元素平均再基于这些计算出微查准率、微查全率、微F1:

$$macro-P=\frac{\overline{TP}}{\overline{TP}+\overline{FP}}$$
$$macro-R=\frac{\overline{TP}}{\overline{TP}+\overline{FN}}$$
$$macro-F1=\frac{2\times macro-P\times macro-R}{macro-P+macro-R}$$

性能评估

ROC与AUC

- 分类阈值 `threshold`

  比如对西瓜及逆袭你给分类的样本预测输出是 [0.0, 1.0]之间的实值

- 截断点 `cut point`

  比如在0.5以下的就是坏瓜，以上的是好瓜。前部分为正例，后半部分为反例

  若重视查准率，截断点靠前；重视查全率，截断点靠后

ROC

受试者工作特征 `Receiver Operating Characteristic` 曲线

- 根据预测结果对样例排序

- 按顺序把样本作为正例进行预测，计算查准率与查全率
- 计算 **真正利率** `True Positive Rate`

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

- 计算 **假正利率** `False Positive Rate`

$$FPR = \frac{FP}{TN + FP}$$

- 以 TPR 为纵轴， FPR 为你横轴绘 ROC 曲线

对于有限的样例可能得不到光滑曲线，则绘制近似曲线

- 给定 m^+ 个正例，m^- 个反例，对预测结果进行排序
- 设置分类阈值最大即所有杨柳为反例，则 TPR = FPR = 0，标记点 （0，0）
- 将分类阈值一次设置为你每个样例的预测值，则每个样例一次为正例，前一个点坐标为（x, y）
  - 若为真正例，则标记 $(x, y + \frac{1}{m^+})$
  - 若为假正例，则标记 $(x + \frac{1}{m^-}, y)$
  - 将两点相连

AUC

`Area Under ROC Curve`，为 ROC 曲线下的面积，用于判断算法的优劣，一般来说包住另一条曲线的更为优化

- AUC = 1，所有正例排在负例前
- AUC = 0，所有负例排在正例前

$$AUC = \frac{1}{2} \sum_{i=1}^{m-1} (x_{i+1} - x_i) \cdot (y_i + y_{i+1})$$

类似于梯形面积公式的计算

代价敏感错误率与代价曲线

许多时候不同类型的错误不能同等对待，比如看病

- 非均等代价 `unequal cost`

  权衡不同类型错误造成的不同损失

- 代价矩阵 `cost matrix`

  $cost_{ij}$: 将第 i 类样本预测为 j 类的代价

  $cost_{ii} = 0$

  以二分类为例

真实类别	预测类别	
	第0类	第1类
第0类	0	cost01
第1类	cost10	0

在非均等代价环境下希望最小化总体代价 `total coast`，以上表格的代价敏感 `cost-sensitive` 错误率为

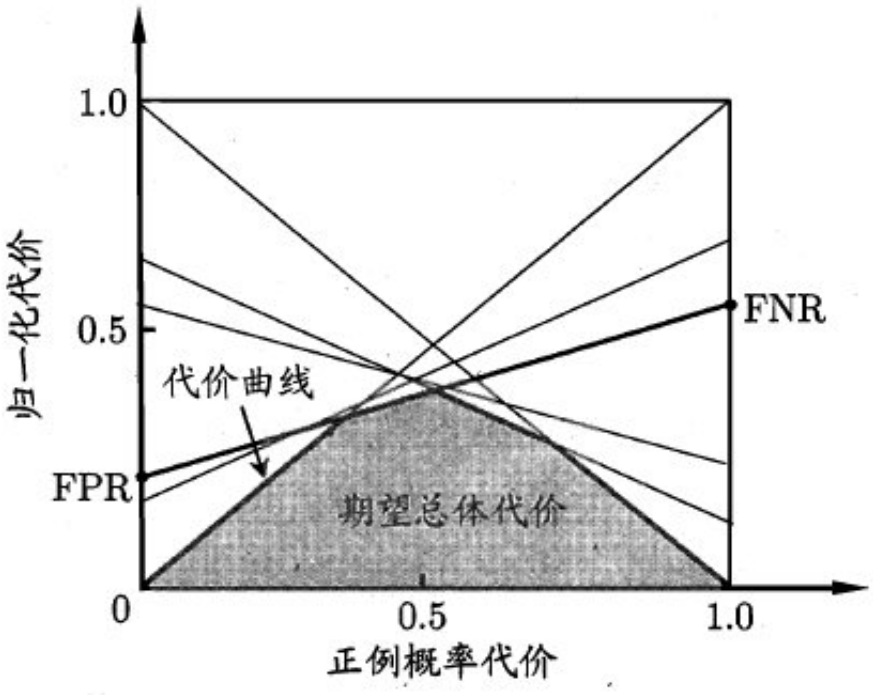
$$E(f;D;cost)=\frac{1}{m}(\sum_{x_i\in D^+}\mathbb{I}(f(x_i\neq y_i)\times cost_{01}+\sum_{x_i\in D^-}\mathbb{I}(f(x_i\neq y_i)\times cost_{10})$$

设 p 为正例概率，纵轴取之归一化为 [0, 1]，则正例概率代价

$$P(+)cost=\frac{p\times cost_{01}}{p\times cost_{01}+(1-p)\times cost_{10}}$$

$$cost_{norm}=\frac{(1-TPR)\times p\times cost_{01}+FPR\times(1-p)\times cost_{10}}{p\times cost_{01}+(1-p)\times cost_{10}}$$

得到期望总体代价图



- 取 ROC 曲线上一点的坐标为 $(TPR\,,\,FPR)$，可计算出 $FNR=1-TPR$
- 在代价平面上绘制一条从 $(0\,,\,FPR)$ 到 $(1\,,\,FNR)$ 的线段，线段下的面积即表示了 *该条件下* 的期望总体代价
- 对于 ROC 曲线上的额每个点重复以上步骤得到所有面积相交的区域面积为即为在 *所有条件下* 学习器的期望总体代价

比较校验

有了评估方法和性能度量，需要对这些评估与度量结果进行比较。**为方面理解默认错误率 ϵ 为性能度量**

假设检验

前提：一个学习器中的错误与其他部分是独立不相关的，所以每次随机采样的错误率分布也是独立不相关的

友情提示：在看之前复习 **概率论**

- 泛化错误率 ϵ

  学习器在一个样本上的犯错概率

- 测试错误率 $\hat{\epsilon}$

  在 m 个测试样本中有 $\hat{\epsilon} \times m$ 个被误分类

- 若泛化错误率为 ϵ 的学习器将样本中的 m' 个样本误分类，则其余样本正确的概率为

$$\epsilon^{m'}\left(1-\epsilon\right)^{(m-m')}$$

- 所以将$\hat{\epsilon} \times m$ 个样本误分类的概率为

$$P(\hat{\epsilon};\epsilon)=\binom{m}{\hat{\epsilon} \times m}\epsilon^{\hat{\epsilon} \times m}(1-\epsilon)^{m-\hat{\epsilon} \times m}$$

求偏导之后可以知道符合 **二项分布**，进行 **二项检验** `Binomial Test`

- 要对泛化错误率是否大于0.3进行检验，书上有个公式

$$\overline{\epsilon}=\max\epsilon \quad s.t. \quad \sum_{i=\epsilon_0 \times m+1}^{m}\binom{m}{i}\epsilon^{i}(1-\epsilon)^{m-i}<\alpha$$

- 该式子的右半部分是一个二项分布公式 $\binom{m}{i}\epsilon^{i}(1-\epsilon)^{m-i}$ ，设

  $\varepsilon_0=0.3$，$H_0:\varepsilon \leqslant \varepsilon_0$，$H_1:\varepsilon>\varepsilon_0$，拒绝域在接受域右侧，此时拒绝域的形式为 $\overline{x} \geqslant k$，拒绝域的边界临界点$|z|=z_{\frac{2}{\alpha}}$，α 通常取0.1，0.05，0.01，0.005

- 样本均值 $\overline{X}=\frac{1}{k}\sum_{i=1}^{k}\hat{\varepsilon_i})$
- 方差 $S^2=\frac{1}{k-1}\sum_{i=1}^{k}(\hat{\varepsilon_i}-\overline{X})^2$
- 得到拒绝域服从 t 分布，当 n 足够大时接近 N(0, 1) 分布

$$\overline{x}=\varepsilon_0+\frac{S^2}{\sqrt{n}}z_\alpha$$

$$z=\frac{\overline{x}-\varepsilon_0}{S/\sqrt{n}} \geqslant z_\alpha=\frac{k-\varepsilon_0}{S/\sqrt{n}}$$

所以如果假设 ε_0 和显著度 α，可计算当错误率为 ε_0 时的临界值 ——在 $1-\alpha$ 概率内能观测到的最大错误率，同时还有左边检验、双边检验等，都是概率论的基础知识。

看到一个图像解释比较清晰的博客https://www.jianshu.com/p/e442814bd370

交叉验证 t 检验

t 检验针对两个正态总体，对于两个学习器 A 和 B，使用 k 折交叉验证得到错误率分别为 ϵ_i^A，ϵ_i^B，同时遵循若两个学习器性能相同，则理论上 $\epsilon_i^A = \epsilon_i^B$

- 对每对结果求差 $\Delta_i = \epsilon_i^A - \epsilon_i^B$ 得到 $\Delta_1, \Delta_2, \ldots, \Delta_k$
- 假设 $H_0 : \epsilon^A = \epsilon^B$，$H_1 : \epsilon^A \neq \epsilon^B$，则

$$\frac{|\overline{\Delta} - (\epsilon^A - \epsilon^B)|}{S/\sqrt{k}} \sim t(k-1)$$

依据

$$|t| = |\frac{\overline{\Delta} - 0}{S/\sqrt{k}}|$$

$$= |\frac{\overline{\Delta}}{S/\sqrt{k}}| \geqslant k = t_{\frac{2}{\alpha}}(k-1)$$

求临界点与拒绝域。

5 x 2交叉验证

在交叉验证的过程中会有训练集重叠的现象，导致错误率实际上不相互独立。缓解这一问题 **做5次2折交叉验证**

- 每次二折交叉验证之前打乱数据，做5次交叉验证
- 取第一次的差值结果求平均
- 取所有10对差值求方差

McNemar 检验

针对 **二分类问题**，回顾一下代价矩阵，得到

algorithm B	algorithm A	
	True	False
True	e00	e01
False	e10	e11

若两个学习器性能相同，则有

$$e_{01} = e_{10},\ |e_{01} - e_{10}| \sim N(1, e_{01} + e_{10})$$

$$\tau_{\chi^2} = \frac{(|e_{01} - e_{10}| - 1)^2}{e_{01} + e_{10}}$$

服从自由度为 1 的 χ^2 分布。

Friedman 检验与 Nemenyi 后续检验

交叉验证 t 检验、McNemar 检验都是在一个数据集上比较两个算法，对于一个数据集上的 **多个算法**，可以进行两辆比较，也可以用给予排序的 Friedman 检验。

Friedman 检验

- 假定 D_1，D_2，D_3，D_4四个数据集与算法 A，B，C
- 用留出法或交叉验证法用每个算法在每个数据集上进行测试，结果按照性能由好到坏排序，并赋值 1，2，…，若性能相同，则平分序值
- 若学习器性能相同则平均序值理论相同，在 N 个数据集上使用 k 个算法，r_i 表示第 i 个算法的平均序值，则 r_i 服从正态分布

$$N \sim (\frac{k+1}{2}, \frac{(k-1)(k+1)}{12})$$

$$\tau_{\chi^2} = \frac{k-1}{k} \cdot \frac{12N}{k^2-1} \sum_{i=1}^{k} (r_i - \frac{k+1}{2})^2$$

$$= \frac{12N}{k(k+1)} (\sum_{i=1}^{k} r_i^2 - \frac{k(k+1)^2}{4})$$

当 k 与 N 都较大时服从自由度为 k - 1 的 χ^2 分布

$$\tau_F = \frac{(N-1)\tau_{\chi^2}}{N(k-1) - \tau_{\chi^2}}$$

τ_F 服从自由度为 k - 1 和 (k - 1)(N - 1) 的 F 分布。

F 检验常用临界值表

$\alpha = 0.05$									
数据集	算法个数 k								
个数 N	2	3	4	5	6	7	8	9	10
4	10.128	5.143	3.863	3.259	2.901	2.661	2.488	2.355	2.250
5	7.709	4.459	3.490	3.007	2.711	2.508	2.359	2.244	2.153
8	5.591	3.739	3.072	2.714	2.485	2.324	2.203	2.109	2.032
10	5.117	3.555	2.960	2.634	2.422	2.272	2.159	2.070	1.998
15	4.600	3.340	2.827	2.537	2.346	2.209	2.104	2.022	1.955
20	4.381	3.245	2.766	2.492	2.310	2.179	2.079	2.000	1.935
$\alpha = 0.1$									
数据集	算法个数 k								
个数 N	2	3	4	5	6	7	8	9	10
4	5.538	3.463	2.813	2.480	2.273	2.130	2.023	1.940	1.874
5	4.545	3.113	2.606	2.333	2.158	2.035	1.943	1.870	1.811
8	3.589	2.726	2.365	2.157	2.019	1.919	1.843	1.782	1.733
10	3.360	2.624	2.299	2.108	1.980	1.886	1.814	1.757	1.710
15	3.102	2.503	2.219	2.048	1.931	1.845	1.779	1.726	1.682
20	2.990	2.448	2.182	2.020	1.909	1.826	1.762	1.711	1.668

Nemenyi 后续检验

当"算法性能相同"假设被拒绝说明算法性能显著不同，这是需要进行 **后续检验** `post-hoc test` 来进一步区分各算法。

Nemenyi 后续检验计算出平均序值差别的临界值域，若两个算法的平均序值差超出了临界值域CD，则相应的置信度 1-α 拒绝假设。

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}}$$

常用 q_α 值表

α	算法个数 k								
	2	3	4	5	6	7	8	9	10
0.05	1.960	2.344	2.569	2.728	2.850	2.949	3.031	3.102	3.164
0.1	1.645	2.052	2.291	2.459	2.589	2.693	2.780	2.855	2.920

偏差与方差

偏差-方差分解 `bias-variance decomposition` 对学习算法泛化性能进行解释——对泛化错误进行拆解

- 测试样本 x 标记 y，在 D 上的标记 y_D ， $f(x;D)$ 为训练集上学习的模型 f 在 x 上的预测输出
- 学习算法的期望预测

$$\overline{f}(x) = \mathbb{E}_D[f(x;D)]$$

- 相同样本数不同训练集产生的方差

$$var(x) = \mathbb{E}_D[(f(x;D) - \overline{f}(x))^2]$$

- 噪声

$$\varepsilon^2 = \mathbb{E}_D[(y_D - y)^2]$$

- 期望结果与真实值的偏差 `bias`

$$bias^2(x) = (\overline{f}(x) - y)^2$$

- 对期望泛化误差分解

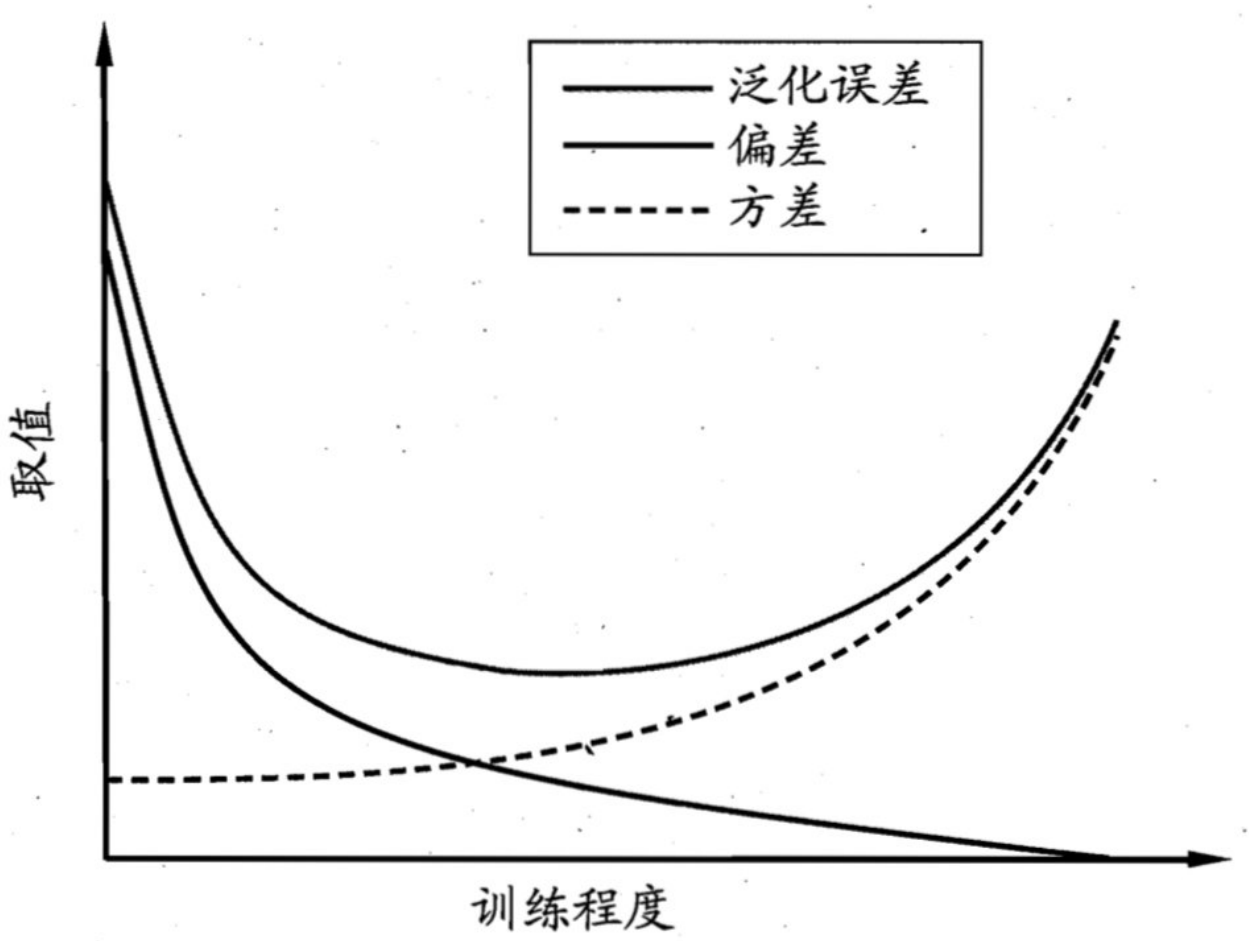
$$\text{假定 } \mathbb{E}_D[y_D - y] = 0$$
$$E(f;D) = \mathbb{E}_D[(f(x;D) - y_D)^2]$$
$$= \mathbb{E}_D[(f(x;D) - \overline{f}(x) + \overline{f}(x) - y_D)^2]$$
$$= \ldots = bias^2(x) + var(x) + \varepsilon^2$$

  泛化误差 = 偏差 + 方差 + 噪声

偏差-方差窘境

偏差与方差一般来说时相矛盾的，**偏差体现学习器预测准确度，方差体现学习器预测稳定性**。

训练程度不断提高的同时，期望预测值与真实值之间偏差越来越小，但算法对数据集波动越来越敏感即方差越来越大。



习题

2.1

数据集包含1000个样本，其中500个正例，500个反例，将其划分为包含70%样本的训练集和30%样本的测试集用于留出法评估，试估算共有多少种划分方式。

$C_{500}^{150} \times C_{500}^{150}$

2.2

数据集包含100个样本，其中正反例各一半，假定学习算法所产生的模型是将新样本预测为训练样本数较多的类别（训练样本数相同时进行随机猜测），试给出用10折交叉验证法和留一法分别对错误率进行评估所得的结果。

十折交叉验证：100个样本分为10组，如果平均分布，取其中1组为测试样本，其余9组为测试训练样本，很显然如果判断为正例则由一半的反例存在，反之亦然，所以错误率为50%；

留一法：100个样本取一个为测试样本，若为正例则剩下的样本中反例多，判断错误，反之亦然，所以为100%。

2.3

若学习器A的F1值比学习器B高，试析A的BEP值是否也比B高。

$$F1 = \frac{2 \times P \times R}{P + R}$$

BEP 为 P-R曲线中查全率=查准率的平衡点，则 P = R

$$\frac{1}{F1} = \frac{P+R}{2 \times P \times R} = \frac{1}{2}(\frac{1}{P} + \frac{1}{R}) = \frac{1}{P} = \frac{1}{R}$$
$$F1^A > F1^B \Rightarrow P^A > P^B,\ R^A > R^B$$

所以 $BEP^A > BEP^B$

2.4

试述真正例率(TPR)、假正例率(FPR)与查准率(P)、查全率(R)之间的联系

$$P = \frac{TP}{TP + FP}$$
$$R = \frac{TP}{TP + FN} = TPR = \frac{TP}{TP + FN}$$
$$FPR = \frac{FP}{TN + FP}$$

P	查准率、准确率	**预测正例**中**真实正例**的比例
R	查全率、召回率	**真实正例**中**预测正例**的比例
TPR	真正利率	**真实正例**中**预测正例**的比例
FPR	假正利率	**真实反例**中**预测正例**的比例

2.5

证明 $AUC = 1 - l_{rank}$ 详情见南瓜书第二章

https://datawhalechina.github.io/pumpkin-book/#/chapter2/chapter2?id=_221

2.6

错误率与 ROC 曲线的关系

RCO 曲线的横纵坐标为 TPR，FPR，所以曲线上的每一个点都可以计算出错误率。

2.7

试证明任意一条ROC曲线都有一条代价曲线与之对应，反之亦然

证明：肯定有一条曲线 (0, FPR)-(1, FNR)

2.8

不是很理解为什么要出一个规范化的题目，找了一个网上的答案，来自https://blog.csdn.net/Snoopy_Yuan/article/details/62240320

Max-min	z-score
简单	计算量大
容易受高杠杆点和离群点影响	对离群点敏感度相对低一些
当加入新值超出当前最大最小范围时重新计算所有之前的结果	每加入新值都要重新计算所有之前结果

之后的几个问题不多赘述