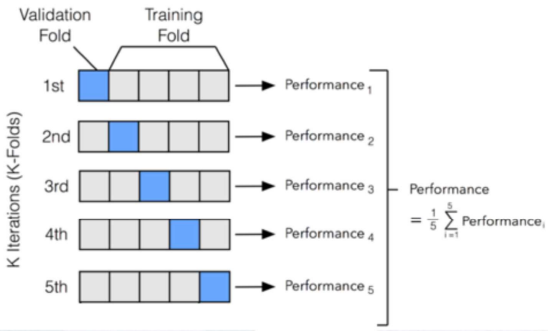
训练过程的特性

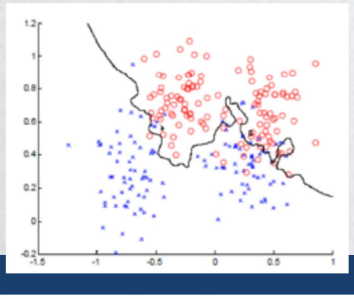
- 随着K增加
  - 分类边界面越来越平滑
  - 训练误差增加
- K是参数，K的选着可以用交叉验证的办法
  - 把训练集进一步分为训练集和验证集
  - 通过验证集的损失误差选择K



40

K近邻总结

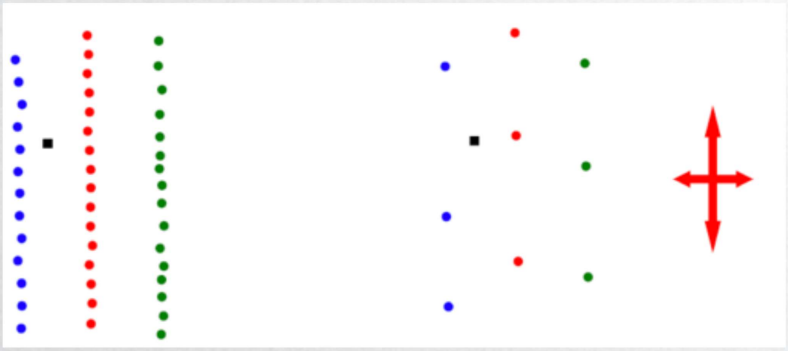
- 优点
  - K近邻分类方法简单有效
  - 可应用到多分类问题上
  - 分类面是非线性的
  - 随着样本数量增加精度会自然提升
  - 只有一个参数K，容易通过交叉验证设置该参数



42

K近邻总结

- 缺点
  - 怎么度量最近，需要特定的距离公式
  - 需要存储和搜索数据集。但可以用索引。

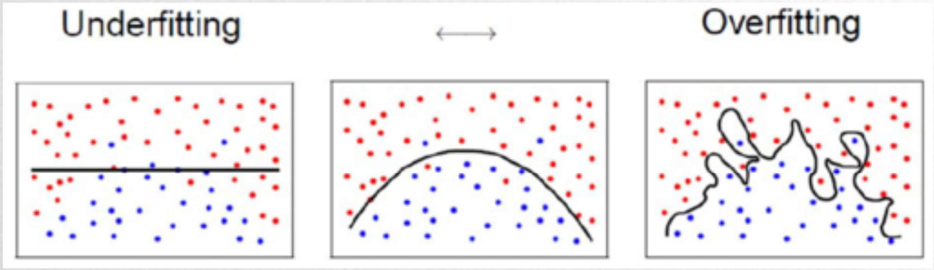


如何设定参数

- 使用验证集
- 把整个数据分成3部分：x
  - 训练集：学习模型参数
  - 验证集：不用来学习，用来选择模型或者调节重要参数
  - 测试集：用来评估模型预测效果，通常评估效果会比验证集差一些
- 我们可以再重新划分数据集，得到新的无偏的预测效果评估。

欠拟合与过拟合

- **Everything Should Be Made as Simple as Possible, But Not Simpler ----Albert Einstein**



精度（误差）的评测效果总是好吗？

- 使用测试集评测分类算法，结果如下：

预测结果

标准答案		正例	负例
	正例	8	12
	负例	8	72

$$\text{Accuracy}=\frac{80}{100}=80\%$$

看指标很不错
但总感觉哪里不对

- 几个关键概念
  - True Positive (TP): 正确识别
  - False Positive (FP): 错误识别
  - True Negative (TN): 正确拒绝
  - False Negative (FN): 错误拒绝

False Positive 错的识别成对的
True Negative 错的识别成错的
False Negative 对的识别成错的

将TP/FP/TN/FN填到下表

标准答案		正例	负例
	正例	TP	FN
	负例	FP	TN

预测结果

精度（误差）的评测效果总是好吗？

- 使用测试集评测分类算法，结果如下：

预测结果

标准答案		正例	负例
	正例	8	12
	负例	8	72

$$\text{Accuracy}=\frac{80}{100}=80\%$$

看指标很不错

但总感觉哪里不对

- 几个关键概念
  - True Positive (TP): 正确识别
  - False Positive (FP): 错误识别
  - True Negative (TN): 正确拒绝
  - False Negative (FN): 错误拒绝

将TP/FP/TN/FN填到下表

标准答案		正例	负例
	正例	TP	FN
	负例	FP	TN

预测结果

精度（误差）的评测效果总是好吗？

- 使用测试集评测分类算法，结果如下：

False Positive：错的识别为对的

预测结果

标准答案		正例	负例
	正例	8	12
	负例	8	72

$$\text{Accuracy}=\frac{80}{100}=80\%$$

看指标很不错

但总感觉哪里不对

- 引入新的指标
  - Precision – 准确率

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

  - Recall – 召回率

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

FN：对的识别成错的

  - F1 Score - F值

$$F1=\frac{2}{\frac{1}{Precision}+\frac{1}{Recall}}$$

请计算左表

精度（误差）的评测效果总是好吗？

- 使用测试集评测分类算法，结果如下：

预测结果

标准答案		正例	负例
	正例	8	12
	负例	8	72

$$\text{Accuracy}=\frac{80}{100}=80\%$$

看指标很不错

但总感觉哪里不对

- 引入新的指标
  - Precision = 50%
  - Recall = 40%
  - F1 Score = 44%

实际分类效果很差！

小节：如何解决一个分类问题

- 提出问题（Question）
- 准备数据（Input Data）
- 选择特征（Features）
- 学习算法（Algorithm）

- 一般在实践中

Q > D > F > A