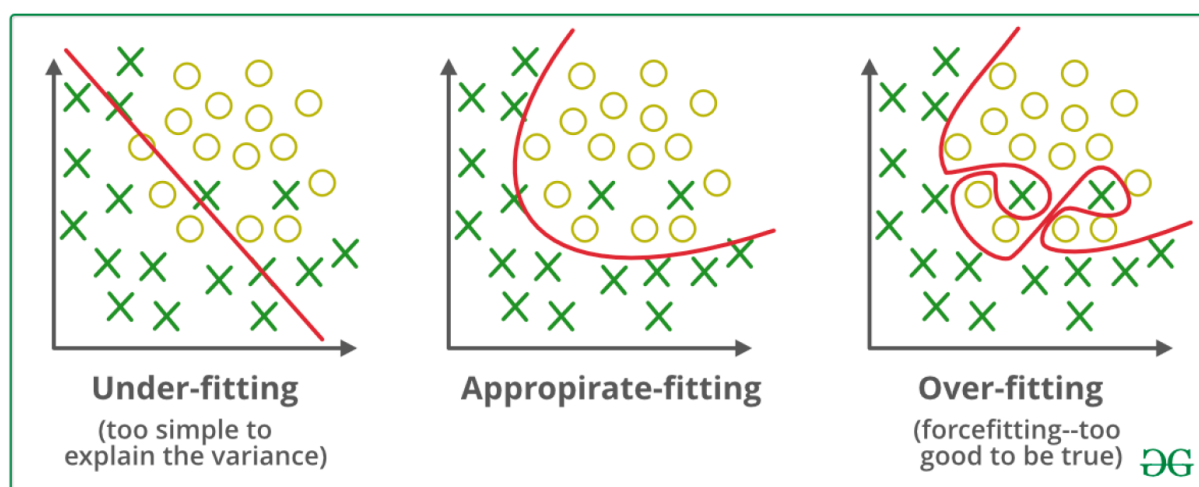




# 机器学习评估



由于引入噪声、模型过于复杂、模型开始记忆训练集，模型很可能过拟合。通常有以下解决方案：

- 用大量的数据训练模型，是模型更加泛化
- 使用K-fold Cross-validation交叉验证
- 使用正则化的工具

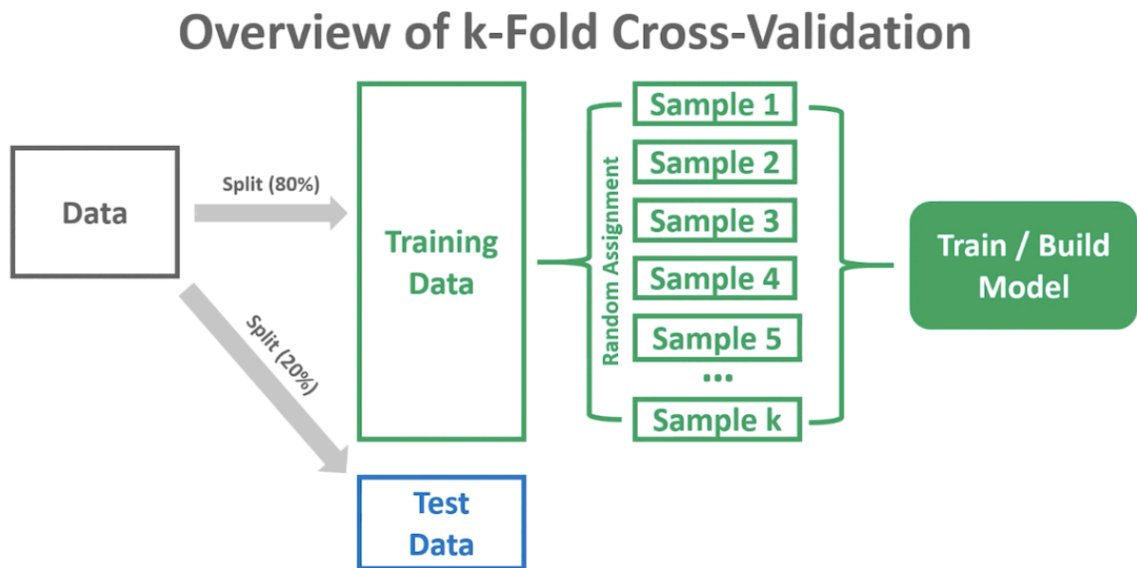
数据分类：

- 训练集：训练模型
- 评估集：帮助调整模型超参数来增强表现

- 测试集：评估模型

### 交叉验证：

- K-Fold Cross-Validation：将数据分成K个块，选取一个块作为验证集，其它作为训练集，并重复选取调参



- Leave One Out Cross Validation (LOOCV)：每次只把训练集中的一个数据作为验证集，对高计算量的小数据集有效，可能很慢，但评估更加准确

### 评估：

- **准确率**：Accuracy = correct/total
  - 局限性：数据是有偏的，比如90%的数据都是Positive
- **Confusion Matrix**：总结模型的True Positive, True Negative, False Positive, False Negative的所有表现
  - Precision：对于Positive情况的精确度，当仅考虑预测正确样本的准确度时使用（比如谷歌不管拒绝了多少潜在好员工，只要招到都是好员工就行）
  - Recall/Sensitivity：代表能够多好地找到所有本身的正确的样本（比如识别每一个病原体）

		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) <b>Type II Error</b>	<b>Sensitivity</b> $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) <b>Type I Error</b>	True Negative (TN)	<b>Specificity</b> $\frac{TN}{(TN + FP)}$
		<b>Precision</b> $\frac{TP}{(TP + FP)}$	<b>Negative Predictive</b>	<b>Accuracy</b> $\frac{TP + TN}{(TP + TN + FP + FN)}$

- Logloss：评估预测模型和真实类别概率之间的一致性，越小的Logloss表现越好

- N=total number of data points
- M=outcomes or classes (usually 0 and 1).
- $X_{ij}$ → whether the actual outcome j is the correct one for data point i.
- $P_{ij}$ → predicted probability that data point i belongs to class j.

Log loss: A measure of accuracy that penalizes overconfidence.

$$LogLoss = \frac{-1}{N} \sum_{i=1}^N \sum_{j=1}^M x_{ij} * \log(p_{ij})$$

- F1-Score：想要Precision和Recall同时表现好

$$\begin{aligned}
 F1 \text{ Score} &= \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} \\
 &= \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
 \end{aligned}$$