COMP 9414 Week 5 - Planning and Reasoning About Action

Author: Junji Duan

Last Update: 2025/10/19

1 W5A: Uncertainty

1.1 Uncertainty

Def: Uncertainty or incertitude refers to situations involving imperfect or unknown information.

Probability gives a way of summarizing this uncertainty.

1.2 Probability Theory

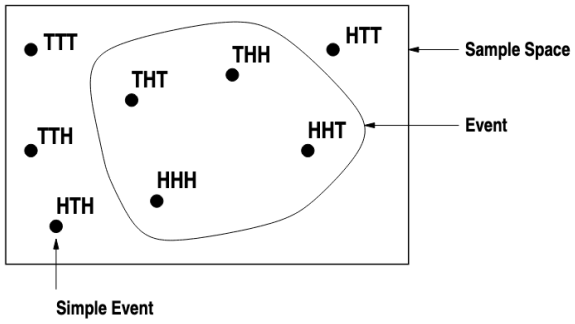
1.2.1 Sample Space and Events

Flip a coin three times, the set of all possible outcomes

$$S = \{\text{TTT}, \text{TTH}, \text{THT}, \text{THH}, \text{HTT}, \text{HTH}, \text{HHT}, \text{HHH}\}$$

Any subset of the sample space is known as an event(事件).

Any singleton subset of the sample space is known as a simple event(基本事件).



1.2.2 Prior Probability 先验概率

$P(A)$ is the prior or unconditional probability that an event A occurs. For example, P(Headache)=0.3

1.2.3 Random Variables 随机变量

Propositions are random variables that can take on several values

$$P(\text{ Weather } = \text{ Sunny }) = 0.8$$
$$P(\text{ Weather } = \text{ Rain }) = 0.2$$

Every random variable X has a domain(域) of possible values $\langle x_1, x_2, \cdots, x_n \rangle$.

Probabilities of all possible values $\mathbf{P}(\text{ Weather }) = \langle 0.8, 0.2 \rangle$ is a probability distribution.

1.2.4 Axioms of Probability

$0 \leq P(A) \leq 1$

$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

$P(A \vee \neg A) = P(A) + P(\neg A) - P(A \wedge \neg A)$

1.3 Conditional Probability and Bayes' Rule

1.3.1 Conditional Probability 条件概率

Def: Conditional probability is a measure of the probability of an event occurring, given that another event. (事件 B 发生的条件下, 事件 A 发生的概率)

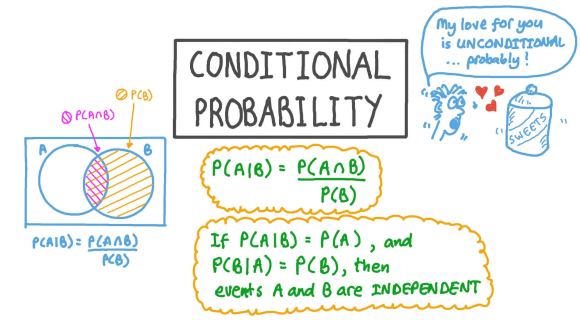
Use conditional or posterior probability $P(A \mid B)$ is the probability of A given that all we know is B. (A 在 B 发生的条件下发生的概率)

Formula: $P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$ provided $P(B) > 0$

Product Rule: $P(A \wedge B) = P(A \mid B) \cdot P(B)$

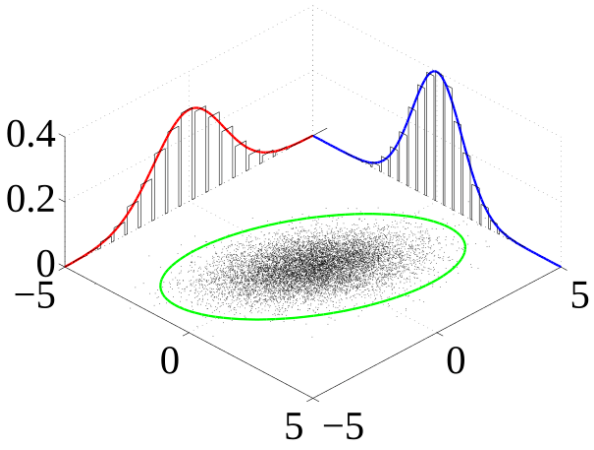
Conditional Probability:
$\mathbf{P}(X \mid Y) = P\left(X = x_i \mid Y = y_j\right)$ for all i, j

Joint Probability (联合概率: 两个事件共同发生的概率): $\mathbf{P}(X, Y) = \mathbf{P}(X \mid Y).\mathbf{P}(Y)$ - a set of equations



1.3.2 Joint Probability Distribution 联合概率分布

Def: A joint probability is the chance that two or more events will happen at the same time. For a joint probability to work, both events must be independent of one another.



Example from Wikipedia:

Rolling a die [edit]

Consider the roll of a fair die and let $A = 1$ if the number is even (i.e. 2, 4, or 6) and $A = 0$ otherwise. Furthermore, let $B = 1$ if the number is prime (i.e. 2, 3, or 5) and $B = 0$ otherwise.

	1	2	3	4	5	6
A	0	1	0	1	0	1
B	0	1	1	0	1	0

Then, the joint distribution of A and B, expressed as a probability mass function, is

$$\mathrm{P}(A=0, B=0) = P\{1\} = \frac{1}{6}, \quad \mathrm{P}(A=1, B=0) = P\{4,6\} = \frac{2}{6},$$
$$\mathrm{P}(A=0, B=1) = P\{3,5\} = \frac{2}{6}, \quad \mathrm{P}(A=1, B=1) = P\{2\} = \frac{1}{6}.$$

These probabilities necessarily sum to 1, since the probability of *some* combination of A and B occurring is 1.

1.3.3 Bayes' Rule 贝叶斯定理

AI systems abandon joint probabilities and work directly with conditional probabilities using Bayes' Rule.

$$P(B \mid A) = \frac{P(A|B)P(B)}{P(A)} \text{ if } P(A) \neq 0$$

Note: If $P(A) = 0, P(B \mid A)$ is undefined.

1.3.4 Conditional Independence 条件独立

An event X is independent of event Y, conditional on background knowledge K, if knowing Y **does not affect** the conditional probability of X given K :

$$P(X \mid K) = P(X \mid Y, K)$$

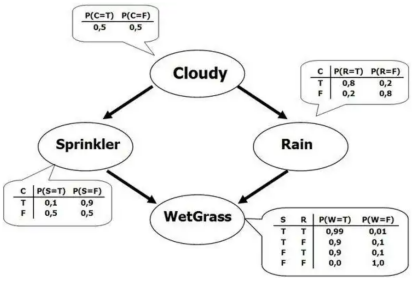
1.4 Bayesian Networks 贝叶斯网络

A Bayesian network (also Bayesian Belief Network, probabilistic network, causal network, knowledge map) is a directed acyclic graph (DAG) where

- Each node corresponds to a random variable
- Directed links connect pairs of nodes - a directed link from node X to node Y means that X has a direct influence on Y
- Each node has a conditional probability table quantifying effect of parents on node

Independence assumption of Bayesian networks

- Each random variable is (conditionally) independent of its nondescendants(非后代) given its parents



Bayesian Network

1.4.1 Semantics of Bayesian Networks

Bayesian network provides a complete description of the domain.

Joint probability distribution can be determined from the network

$$P\left(X_1, X_2, \cdots, X_n\right) = \prod_{i=1}^{n} P\left(X_i \mid \text{ Parents } \left(X_i\right)\right)$$

- 变量：R=Rain（下雨），S=Sprinkler（洒水器），W=WetGrass（草地湿）
- 依赖：S 由 R 决定；W 由 S, R 决定；R 为根节点
- 取值（示例）：
  $P(R) = 0.20,\ P(\neg R) = 0.80$
  $P(S \mid R) = 0.01,\ P(S \mid \neg R) = 0.40$
  $P(W \mid S, R) = 0.99,\ P(W \mid S, \neg R) = 0.90,\ P(W \mid \neg S, R) = 0.80,\ P(W \mid \neg S, \neg R) = 0.01$

联合分布可分解：

$$P(X_1, \ldots, X_n) = \prod_i P(X_i \mid \text{Parents}(X_i))$$

例如，$P(W \wedge S \wedge \neg R) =$

$$P(W \mid S \wedge \neg R) \cdot P(S \mid \neg R) \cdot P(\neg R) =$$
$$0.90 \times 0.40 \times 0.80 = 0.288$$

Bayesian network is a complete and non-redundant representation of domain (and can be far more compact than joint probability distribution)

Chain Rule: Use conditional probabilities to decompose conjunctions

$$P(X_1 \wedge \cdots \wedge X_n) = P(X_1) \cdot P(X_2 \mid X_1) \cdot P(X_3 \mid X_1 \wedge X_2) \cdot \cdots \cdot P(X_n \mid X_1 \wedge \cdots \wedge X_{n-1}) \tag{1}$$

Notice:

$$P(X_i \mid X_1 \wedge X_2 \wedge \cdots \wedge X_{i-1}) = P(X_i \mid \pi_{X_i})$$

$$P(X_1, X_2, \cdots, X_n) = \prod_{i=1}^{n} P(X_i \mid \pi_{X_i})$$

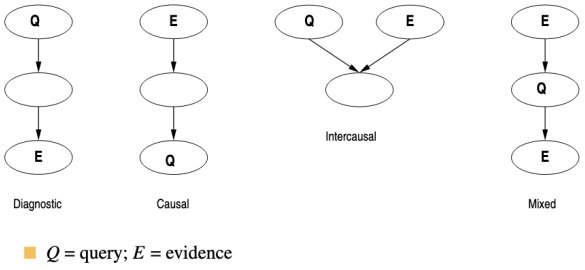
1.4.2 Inference in Bayesian Networks

Diagnostic Inference(诊断推理): From effects to causes(从结果推原因)
$P($ Flu $\mid$ JohnCalls $) = 0.016$

Causal Inference(因果推理): From causes to effects(从原因推结果)
$P($ JohnCalls $\mid$ Burglary $) = 0.85; P($ MaryCalls $\mid$ Burglary $) = 0.67$

Intercausal Inference(因果间推理): Explaining away
$P($ Burglary $\mid$ Alarm $) = 0.3736$ but adding evidence, $P($ Burglary $\mid$ Alarm $\wedge$ Earthquake $) = 0.003$; despite the fact that burglaries and earthquakes are independent, the presence of one makes the other much less likely

Mixed Inference(混合推理): Combinations of the patterns above
Diagnostic + Causal: P(Alarm | JohnCalls $\wedge\neg$ Earthquake)
Intercausal + Diagnostic: P(Burglary|JohnCalls $\wedge\neg$ Earthquake)



■ Q = query; E = evidence

1.4.3 Facts: Calculation using Bayesian Networks

Fact 1: Consider random variable X with parents $Y_1, Y_2, \cdots, Y_n$

$$P(X \mid Y_1 \wedge \cdots \wedge Y_n \wedge Z) = P(X \mid Y_1 \wedge \cdots \wedge Y_n)$$

if Z doesn't involve a descendant of X (including X itself)

Fact 2: If $Y_1, \cdots, Y_n$ are pairwise disjoint and exhaust all possibilities

$$P(X) = \Sigma P(X \wedge Y_i) = \Sigma P(X \mid Y_i) \cdot P(Y_i)$$
$$P(X \mid Z) = \Sigma P(X \wedge Y_i \mid Z)$$

Fact 3: $P(X \mid Z) = P(X \mid Y \wedge Z).P(Y \mid Z) + P(X \mid \neg Y \wedge Z).P(\neg Y \mid Z)$, since $X \wedge Z \equiv (X \wedge Y \wedge Z) \vee (X \wedge \neg Y \wedge Z)$ (conditional version of Fact 2)

Fact 4: $P(X \wedge Y) = P(X) * P(Y)$ if X, Y are independent

2 W5B: Learning

2.1 Machine Learning

2.1.1 Types of Learning

Supervised Learning 监督学习

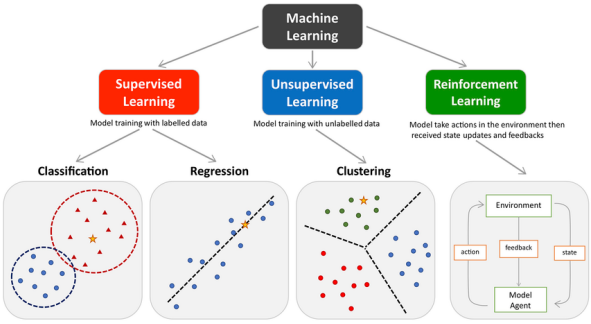
- Agent is presented with examples of inputs and their target outputs, and must learn a function from inputs to outputs that agrees with the training examples and generalizes to new examples

Unsupervised Learning 无监督学习

- Agent is only presented with a series of inputs, and must find useful patterns in these inputs

Reinforcement Learning 强化学习

- Agent is not presented with target outputs for each input, but is periodically given a reward, and must learn to maximize (expected) rewards over time



2.2 Supervised Learning

Given training set and test sets of items, with each training item labeled by a feature vector and target output.

Learner must learn a model that can predict the target output for any given item (characterized by its set of features).

In supervised learning, the learner builds the model using only the features and labels of the training examples—whether presented in batch or online and in random or time order—and must not use the test set during modeling or tuning(调参).

Model is evaluated by its performance on predicting the output for each item in the test set.

2.2.1 Methods vs Models

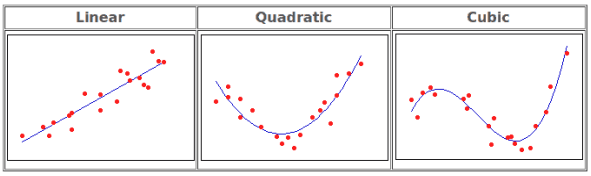
Various learning methods can be used to generate models: Decision Trees, Support Vector Machines, Neural Networks/Deep Learning

In practice, methods are judged via model performance on limited benchmarks(基准), but dataset scarcity, sensitivity to problem setup and hyperparameters, and users' focus on deployable models create a gap between "evaluating methods" and "delivering models."

2.2.2 Supervised Learning –Methodology

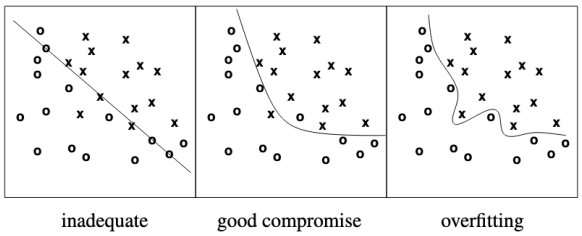
In supervised learning, you engineer and represent features, preprocess raw data, select candidate models and training settings, then evaluate rigorously against baselines with proper validation and expert sanity(reasonable and sensible behaviour or thinking) checks.

2.2.3 Curve Fitting



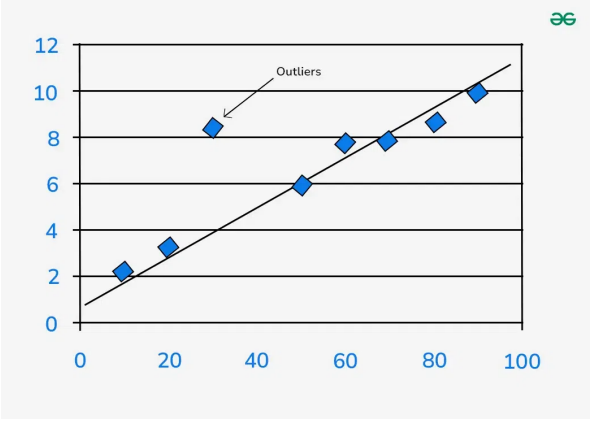
2.2.4 Ockham's Razor

The most likely hypothesis is the simplest one consistent with the data.

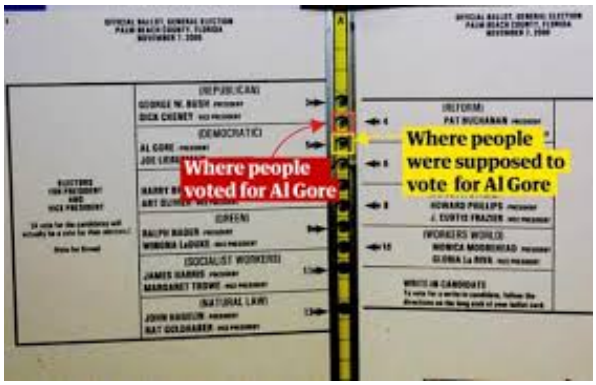


Since there can be noise in the measurements, in practice need to make a tradeoff between simplicity of the hypothesis and how well it fits the data.

2.2.5 Outliers

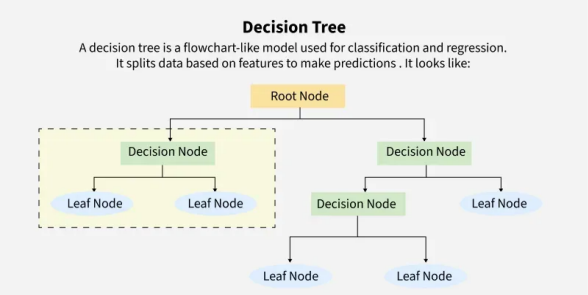


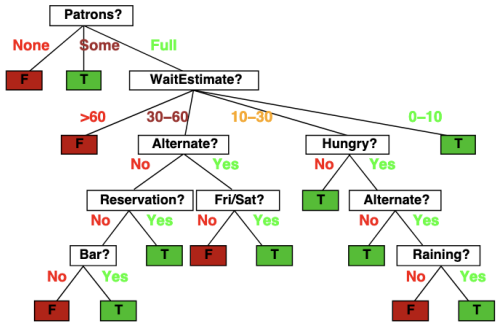
2.2.6 Butterfly Ballot



The Butterfly Ballot shows that small, systematic flaws in how inputs are presented and collected can dominate outcomes—so machine learning must co-optimize model, features, UI/flow, accessibility, causal validation, and live monitoring as a single, end-to-end system.

2.2.7 Decision Tree Learning





Provided the training set is not inconsistent, attributes can be split in any order to produce a tree that correctly classifies all examples in the training set.

However, what is needed is a tree likely to generalize –correctly classify the (unseen) examples in the test set.

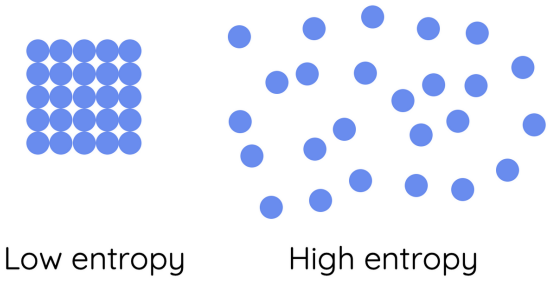
In view of Ockham’ s Razor, a simpler hypothesis is preferred –“simpler” = smaller tree.

2.2.8 Pruning 1: Choosing an Attribute to Split

An attribute is “more informative” if—like Patrons—it splits the data into purer (nearly all-positive or all-negative) subsets, a quality measured by lower entropy, so a parsimonious tree is built by greedily choosing the split that minimizes entropy at each step.

即: 如果一个属性（如 Patrons）能把数据切成更“纯”的子集（几乎全正或全负），它就更有信息量；这种信息量用熵来度量，因此每一步选择使熵最小的划分就能得到更简洁的决策树。

2.2.9 Pruning 1: Entropy and Huffmann Coding



Entropy is a measure of “randomness” (lack of uniformity).

Split based on information gain

- Loss of entropy based on “communicating” value of attribute (选那个能带来“熵下降”最多的属性来划分)

公式总结:

- **直观**：熵衡量**不确定性/随机性**。分布越平均（更难预测），熵越高；越偏向某类（更好预测），熵越低。
- **公式**：对离散分布 $p_1, \ldots, p_n$，

$$H(p_1, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log_2 p_i \quad \text{（单位：bits）}$$

例：$H(0.5, 0.5) = 1$ bit；$H(0.5, 0.25, 0.25) = 1.5$ bits。

- **对分类节点**：若该节点有 p 个正例、n 个反例，则

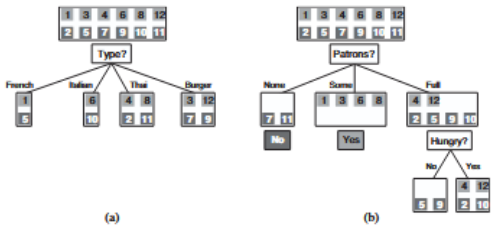
$$H = H\left(\frac{p}{p+n}, \frac{n}{p+n}\right)$$

例如餐馆数据集中父节点 $p = n = 6 \Rightarrow H = 1$ bit（极难预测：一半一半）。

2.2.10 Pruning 1: Information Gain

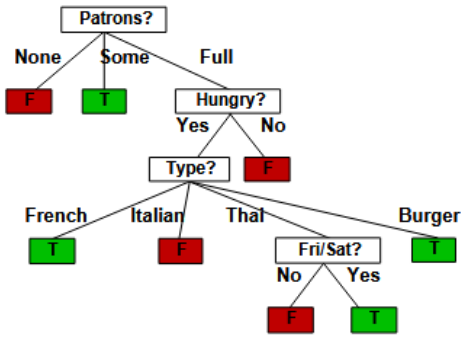
$$\begin{aligned}\text{For Patrons, Entropy} &= \frac{1}{6}(0)+\frac{1}{3}(0)+\frac{1}{2}[-\frac{1}{3}\log_2(\frac{1}{3})-\frac{2}{3}\log_2(\frac{2}{3})] \\ &= 0+0+\frac{1}{2}[\frac{1}{3}(1.585)+\frac{2}{3}(0.585)] = 0.459 \\ \text{For Type, Entropy} &= \frac{1}{6}(1)+\frac{1}{6}(1)+\frac{1}{3}(1)+\frac{1}{3}(1) = 1\end{aligned}$$

2.2.11 Pruning 2: Choosing Next Attribute



After splitting on Patrons, split the node Patrons=Full on Hungry

Induced Decision Tree（Final version）



2.2.12 Laplace Error and Pruning

Laplace Error（叶子）：当把该叶子里的多数类作为预测时，对未见测试样本的预期错误率 (Laplace error)

$$E = 1 - \frac{n+1}{N+k}$$

其中 N = 该节点训练样本总数，n = 多数类 (majority class) 样本数，k = 类别数（平滑：法则 of succession）。

剪枝准则：若子节点的加权平均 Laplace 误差 > 父节点的 Laplace 误差，则剪掉子节点（把父节点当作叶子）。

- 子树 " 看起来更复杂 "，但对未来样本并没有让预期错误率更低（甚至更高），那就剪去。

Example:

Should the children of this node be pruned or not?
Left child has class frequencies [7,3]

$$E = 1 - \frac{n+1}{N+k} = 1 - \frac{7+1}{10+2} = 0.333$$

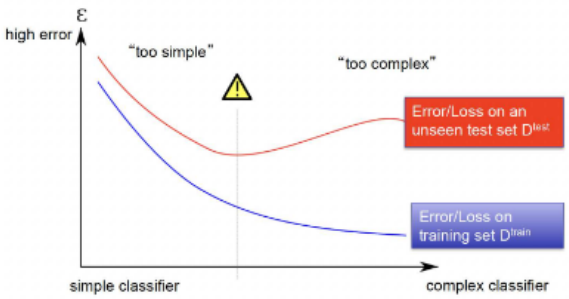
Right child has $E = 0.429$
Parent node has $E = 0.412$
Average for Left and Right child is

$$E = \frac{10}{15}(0.333) + \frac{5}{15}(0.429) = 0.365$$

Since $0.365 < 0.412$, children should not be pruned

2.3 Comparing Methods

2.3.1 Simple vs Complex Models



2.3.2 How to Choose a Method

When we train the same polynomial regressor and repeatedly sample different training sets D, we obtain different models h_D. Some happen to fit well, while others overfit or underfit. The expected error averages over these outcomes to assess the method's average performance in a "repeat-theexperiment" world-which is what we really care about when judging how good the method is.

2.3.3 Bias-Variance Decomposition of Expected Error

Regression setting:

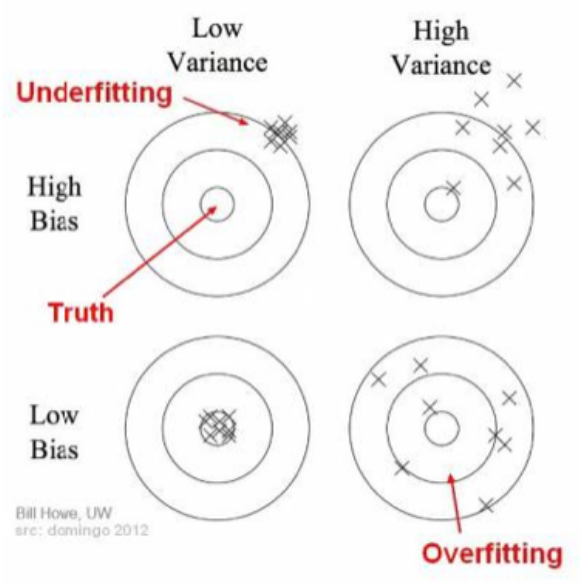
- f : True function that can include noise
- $\bar{f}$: Expected true function = "Average"
- f over datasets D

- h_D : Learned model (hypothesis) from dataset D
- h : Expected model (hypothesis) over datasets D
- h = "Average" model (hypothesis) over learned hypotheses

Result. The expected error $\mathbb{E}_{\{D,x\}}[$ Error$]=$

$$\begin{aligned} & \mathbb{E}_{\{D,x\}}[(\bar{f}(x)-\bar{h}(x))^2] && \text{Bias}^2 \text{ (how far } \bar{h} \text{ is from } \bar{f} \text{ over all } x) \\ + \ & \mathbb{E}_{\{D,x\}}[(h_D(x)-\bar{h}(x))^2] && \text{Variance (how much } h_D \text{ varies around mean } \bar{h}) \\ + \ & \mathbb{E}_{\{D,x\}}[(f(x)-\bar{f}(x))^2] && \text{Noise (irreducible error in the problem)} \end{aligned}$$

2.3.4 High vs Low Bias and Variance

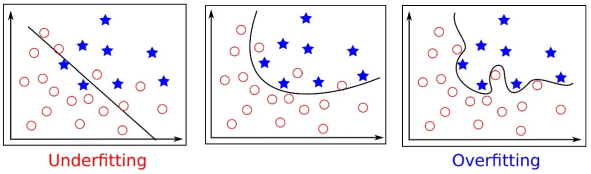


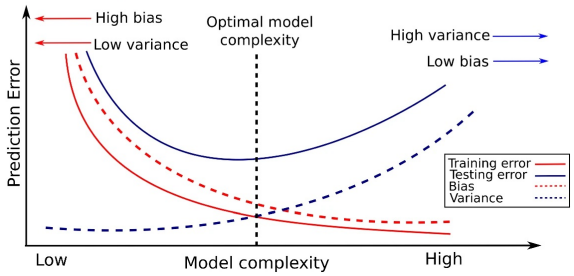
Each plot shows errors using one method to generate different models (hypotheses) h_D for different datasets D, over all instances x

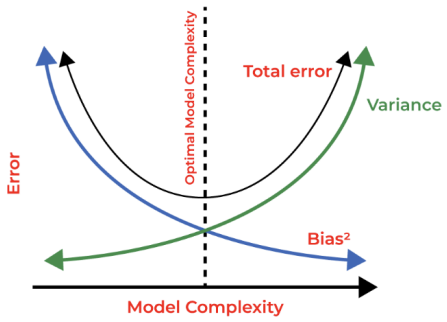
2.3.5 Underfitting vs Overfitting

- High Bias
  - Sign of underfitting: model cannot represent the true function
  - Choose more complex hypothesis space
- High Variance
  - Sign of overfitting: model captures the noise in the data
  - Choose less complex hypothesis space
- Decision Trees prone to overfitting
- Pruning aims to reduce variance (less complex models)
  - ... without increasing bias

2.3.6 Bias-Variance Tradeoff Schematically







2.3.7 Ways to Reduce Bias/Variance

- Reduce Bias
  - More complex models (hypothesis space)
  - Use more features (more discrimination)
  - Apply less regularization (smoothing, pruning, tuning, etc.)
  - Use more training data
- Reduce Variance
  - Less complex models (hypothesis space)
  - Use fewer features (reduces noise)
  - Apply more regularization (smoothing, pruning, tuning, etc.)
  - Stop training sooner

参考文献

[1] UNSW COMP9414 Lecture slides –Prof. W.Wobcke

[2] Other online resources (eg.graph)