**Exercise 10-1** Bias-Variance Decomposition

Assume a fixed distribution $P(x)$ over $x$ and a dependent variable $y = f(x) + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma^2)$. We try to model this relationship by a function $\hat{f}(x, w)$ whose parameter $w$ we learn. For short, we write $f(x) = f$ and $\hat{f}(x, w) = \hat{f}$. Consider the expected square loss $\mathbb{E}[L] = \mathbb{E}[(\hat{f} - y)^2]$.

(a) Show that: $\mathbb{E}[(\hat{f} - y)^2] = \mathbb{E}[(\hat{f} - f)^2] + \mathbb{E}[(f - y)^2]$
with $\mathbb{E}[(f - y)^2] = \mathrm{Var}[y] = \sigma^2$ being the intrinsic noise of the data.

Hint: Make use of the calculation rules for the expected value: For two random variables $X$ and $Y$ and a constant $c$ it holds that: $\mathbb{E}[cX + Y] = c\mathbb{E}[X] + \mathbb{E}[Y]$. Moreover, if $X$ and $Y$ are independent: $\mathbb{E}[X \cdot Y] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$.

## Possible Solution

$$\begin{aligned}
\mathbb{E}[(\hat{f} - y)^2] &= \mathbb{E}[(\hat{f} - f - \epsilon)^2] \\
&= \mathbb{E}[(\hat{f} - f)^2 + \epsilon^2 - 2(\hat{f} - f)\epsilon] \\
&= \mathbb{E}[(\hat{f} - f)^2] + \mathbb{E}[\epsilon^2] - 2\mathbb{E}[\hat{f}\epsilon] + 2\mathbb{E}[f\epsilon] \\
&= \mathbb{E}[(\hat{f} - f)^2] + \mathrm{Var}[\epsilon] - 2\mathbb{E}[\hat{f}] \cdot \mathbb{E}[\epsilon] + 2f \cdot \mathbb{E}[\epsilon] && \text{since } \epsilon \perp \hat{f} \text{ and } f \text{ is deterministic} \\
&= \mathbb{E}[(\hat{f} - f)^2] + \sigma^2 && \text{since } \mathbb{E}[\epsilon] = 0
\end{aligned}$$

In the 4th step we used that $\mathbb{E}[\epsilon^2] = \mathrm{Var}[\epsilon] + (\mathbb{E}[\epsilon])^2 = \sigma^2 + 0$.

Also note that $\mathbb{E}[\epsilon^2] = \mathbb{E}[(y - f)^2] = \mathbb{E}[(y - \mathbb{E}[y])^2] = \mathrm{Var}[y]$

(b) Now show that: $\mathbb{E}[(\hat{f} - f)^2] = \mathbb{E}[(\hat{f} - \mathbb{E}[\hat{f}])^2] + (\mathbb{E}[\hat{f}] - f)^2$
where $\mathbb{E}[(\hat{f} - \mathbb{E}[\hat{f}])^2]$ is the variance and $(\mathbb{E}[\hat{f}] - f)^2$ the squared bias of our estimation $\hat{f}$.

Hint: Substract and add $\mathbb{E}[\hat{f}]$ to the squared term in the loss and simplify.

## Possible Solution

$$\begin{aligned}
\mathbb{E}[(\hat{f} - f)^2] &= \mathbb{E}[(\hat{f} - \mathbb{E}[\hat{f}] + \mathbb{E}[\hat{f}] - f)^2] \\
&= \mathbb{E}[(\hat{f} - \mathbb{E}[\hat{f}])^2] + \mathbb{E}[(\mathbb{E}[\hat{f}] - f)^2] + 2\mathbb{E}[(\hat{f} - \mathbb{E}[\hat{f}])(\mathbb{E}[\hat{f}] - f)] \\
&= \mathrm{Var}[\hat{f}] + (\mathbb{E}[\hat{f}] - f)^2 + 2(\mathbb{E}[\hat{f}] - f) \cdot \mathbb{E}[(\hat{f} - \mathbb{E}[\hat{f}])] && \text{since } (\mathbb{E}[\hat{f}] - f) \text{ is deterministic} \\
&= \mathrm{Var}[\hat{f}] + (\mathrm{Bias}(\hat{f}))^2 + 2(\mathbb{E}[\hat{f}] - f) \cdot (\mathbb{E}[\hat{f}] - \mathbb{E}[\hat{f}]) && \text{since } \mathbb{E}[\mathbb{E}[X]] = \mathbb{E}[X] \\
&= \mathrm{Var}[\hat{f}] + (\mathrm{Bias}(\hat{f}))^2
\end{aligned}$$

(c) With that we have:

$$\begin{aligned}
\mathbb{E}[L] = \mathbb{E}[(\hat{f} - y)^2] &= \mathbb{E}[(\hat{f} - \mathbb{E}[\hat{f}])^2] + (\mathbb{E}[\hat{f}] - f)^2 + \mathbb{E}[(f - y)^2] \\
&= \mathrm{Var}[\hat{f}] + (\mathrm{Bias}(\hat{f}))^2 + \sigma^2.
\end{aligned}$$

What is the minimum of the expected loss?

## Possible Solution

For the best possible model we would have $\hat{f}(x) = f(x)$ and thus

$$\mathbb{E}[L] = \mathbb{E}[(f - f)^2] + \mathbb{E}[(f - y)^2] = 0 + \sigma^2 = \sigma^2.$$

In this case our model has zero bias and variance and only the intrinsic noise remains as loss.

**Exercise 10-2** Bayesian Network

Consider the following network, in which a mouse agent is reasoning about the behavior of a cat. The mouse really wants to know whether the cat will attack (A), which depends on whether the cat is hungry (H) and whether the cat is sleepy (S). The mouse can observe two things, whether the cat is sleepy (S) and whether the cat has a collar (C). The cat is more often sleepy (S) when it's either full (f) or starved (v) than when it is peckish (p) and the collar (C) tends to indicate that the cat is not starved. Note that entries are omitted, such as $P(C = \neg c)$, when their complements are given.

$P(C)$

| C | P(C) |
|---|---|
| $c$ | 0.4 |

$P(H \mid C)$

| H | C | P |
|---|---|---|
| f | c | 0.7 |
| v | c | 0.1 |
| p | c | 0.2 |
| f | ¬c | 0.2 |
| v | ¬c | 0.5 |
| p | ¬c | 0.3 |

$P(S \mid H)$

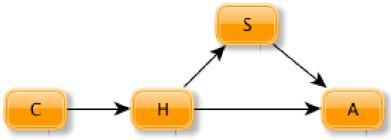
| S | H | P |
|---|---|---|
| s | f | 0.9 |
| s | v | 0.6 |
| s | p | 0.3 |

$P(A \mid H, S)$

| A | H | S | P |
|---|---|---|---|
| a | f | s | 0.01 |
| a | f | ¬s | 0.1 |
| a | v | s | 0.4 |
| a | v | ¬s | 0.9 |
| a | p | s | 0.2 |
| a | p | ¬s | 0.7 |

(a) Draw the Bayesian network corresponding to the above joint probability distribution on $C, H, S, A$.

## Possible Solution



(b) Compute the following probabilities:

- P(A=a, C=c, S=s, H=f)
- P(A=a, C=c, S=s)
- P(C=c, S=s)
- P(A=a | C=c, S=s)

## Possible Solution

- $P(A = a, C = c, S = s, H = f) = .4 * .7 * .9 * .01 = .00252$

- P(A=a, C=c, S=s)

$$P(A = a, C = c, S = s, H = f) + P(A = a, C = c, S = s, H = v) + P(A = a, C = c, S = s, H = p) =$$
$$= (0.4 * 0.7 * 0.9 * 0.01) + (0.4 * 0.1 * 0.6 * 0.4) + (0.4 * 0.2 * 0.3 * 0.2) = 0.01692$$

- P(C=c, S=s)

$$P(C = c, S = s, H = f) + P(C = c, S = s, H = v) + P(C = c, S = s, H = p) =$$
$$(0.4 * 0.7 * 0.9) + (0.4 * 0.1 * 0.6) + (0.4 * 0.2 * 0.3) = 0.3$$

- $P(A = a \mid C = c, S = s) = \frac{P(A=a,C=c,S=s)}{P(C=c,S=s)} = \frac{0.01692}{0.3} = 0.0564$

(c) The mouse is trying to figure out whether it should run out its hole and eat the cheese (E) or do nothing (N). If the mouse hides, nothing happens but it stays hungry. If the mouse runs out to eat the cheese and the cat attacks, the mouse dies (which has a low utility). Otherwise, if the mouse tries to eat the cheese and the cat does not attack, it gets to eat tasty cheese (high utility).

Utilities

| Cat ready to attack (A) | Mouse's action | Utility |
|---|---|---|
| a | E | x |
| ¬ a | E | 5 |
| Any | N | -2 |

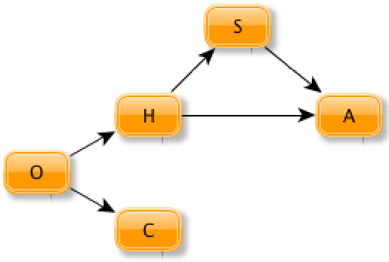
- Suppose in the above table $x$ is $-10$. The mouse sees that the cat has a collar on and is sleepy. What is the utility of trying to eat the cheese? What about doing nothing? Which option should the mouse choose?
- What should the utility of dying (x in the above table) be in order for the mouse to be ambivalent between running for the cheese and doing nothing? Again, the cat is wearing a collar and is sleepy.

### Possible Solution

- The utility of doing nothing is $-2$. The utility of going for the cheese is $P(A = a \mid C = c, S = s) * x + P(A = \neg a \mid C = c, S = s) * 5 = 4.154$. The mouse should go for the cheese.
- We need to find an $x$ such that the utility of going for the food is equal to the utility of doing nothing. Thus, $x$, must solve
$P(A = a \mid C = c, S = s) * x + P(A = \neg a \mid C = c, S = s) * 5 = -2.$
$\rightarrow x = -119.113.$

(d) You may have noticed that one of the variables in the network is “collar“, which according to the CPTs (conditional probability table) causes hunger. However, the real relationship is of correlation, not causation. Introduce a new node $O$ for “owner“ and draw a network which better models the true relationship between the variables. $C$ and $H$ should be independent contitioned on $O$.

### Possible Solution



**Exercise 10-3** Maximum A Posteriori Rule

Consider the lifetime of 2 different types of devices, which we will refer to as type $A$ and type $B$. The lifetime of a type-$A$ device is exponentially distributed with parameter $\lambda$. The lifetime of type-$B$ is exponentially distributed with parameter $\mu$, with $\mu > \lambda > 0$. Assuming you have a box full of devices of the same type, and you would like to know whether they are of type $A$ or $B$. Assume an *a priori* probability of 1/3 that the box contains devices of type B.

(a) You observe the value $t_1$ of the lifetime, $T_1$, of a device. A MAP decision rule implies that the device is of type A if and only if $t_1 > \alpha$. Assuming that $\mu \geq 2\lambda$, find $\alpha$. Express your answer in terms of $\mu$ and $\lambda$.
(*Hint: Exponential distribution is given by* $f(x; \lambda) = \lambda e^{-\lambda x}$ *iff* $x \geq 0$)

(b) Assuming that $\mu \geq 2\lambda$, what is the probability of error of the MAP estimator?

(c) Assume that $\lambda = 2$ and $\mu = 3$. Find the LMS estimate of $T_2$, the lifetime of another device from the same box, based on observing $T_1 = 2$. Assume that conditioned on the device type, the lifetimes of devices are independent.

**Exercise 10-4** Markov Property

For each of the following definitions of the state $X_k$ at time $k$ (for $k = 1, 2, \ldots$), determine whether the Markov property is statisfied by the sequence $X_1, X_2, \ldots$.

A fair six-sided die (with sides labeled 1,2,3,4,5,6) is rolled repeatedly and independently.

(a) Let $X_k$ denote the largest number obtained in the first $k$ rolls. Does the sequence $X_1, X_2 \ldots$ satisfy the Markov property?

(b) Let $X_k$ denote the number of 6's obtained in the first $k$ rolls, up to a maximum of ten. (That is, if ten or more 6's are obtained in the first $k$ rolls, then $X_k = 10$.) Does the sequence $X_1, X_2, \ldots$ satisfy the Markov property?

(c) Let $Y_k$ denote the result of the $k^{th}$ roll. Let $X_1 = Y_1$ and for $k \geq 2$, let $X_k = Y_k + Y_{k-1}$. Does the sequence $X_1, X_2, \ldots$ satisfy the Markov property?

(d) Let $Y_k = 1$ if the $k^{th}$ roll results is an odd number; and $Y_k = 0$ otherwise. Let $X_1 = Y_1$ and for $k \geq 2$ let $X_k = Y_k \cdot X_{k-1}$. Does the sequence $X_1, X_2, \ldots$ satisfy the Markov property?

## Possible Solution

(a) Since the state $X_k$ is the largest number obtained in $k$ rolls, the set of states is $S = \{1,2,3,4,5,6\}$. Given the largest number obtained in the first $k$ rolls, the probability distribution of the largest number obtained in the first $k+1$ rolls no longer depends on what the largest number obtained was in the first $k-1$ rolls (or in the first $k-2$ rolls, etc.). Therefore the Markov property is satisfied.

For $i, j \in \{1,2,3,4,5,6\}$, the transition probabilitiese are:

$$p_{i,j} = \begin{cases} 0, & \text{if} j < i \\ i/6, & \text{if} j = i \\ 1 - i/6, & \text{if} j > i \end{cases}$$

(b) Since the state $X_k$ is the number of 6's in the first $k$ rolls, the set of states is $S = \{0,1,2,\dots,10\}$. The probability of getting a 6 in a given trial is 1/6. Given the number of 6's in the first $k$ rolls, the probabilitiy distribution of the number of 6's in the first $k+1$ rolls no longer dependes on the number of 6's in the first $k-1$ rolls (or in the first $k-2$ rolls, etc.). Therefore the Markov property is satisfied. Thus $p_{10,10} = 1$, and for $i \leq 9$, the transition probabilities are:

$$p_{i,j} = \begin{cases} 1/6, & \text{if} j = i+1 \\ 5/6, & \text{if} j = i \\ 0, & \text{otherwise} \end{cases}$$

(c) We have:

$$P(X_3 = 2 \mid X_2 = 3, X_1 = 1) = P(Y_2 + Y_3 = 2 \mid Y_1 = 1, Y_2 = 2)$$
$$P(Y_3 = 0 \mid Y_1 = 1, Y_2 = 2) = 0$$

but

$$P(X_3 = 2 \mid X_2 = 3, X_1 = 2) = P(Y_2 + Y_3 = 2 \mid Y_1 = 2, Y_2 = 1)$$
$$P(Y_3 = 1 \mid Y_1 = 2, Y_2 = 1) = P(Y_3 = 1) = 1/6$$

and therefore the Markov property is violated.

(d) At each stage, $Y_k$ has equal probability of being 0 or 1. Since $X_k = Y_k \cdot X_{k-1}$, and we assume independent rolls, clearly $X_k$ depends only on the $k^{th}$ roll and the value of $X_{k-1}$. Therefore the Markov property is satisfied.

The transition probabilities are $p_{00} = 1$, $p_{01} = 0$, $p_{10} = \frac{1}{2}$, and $p_{11} = \frac{1}{2}$