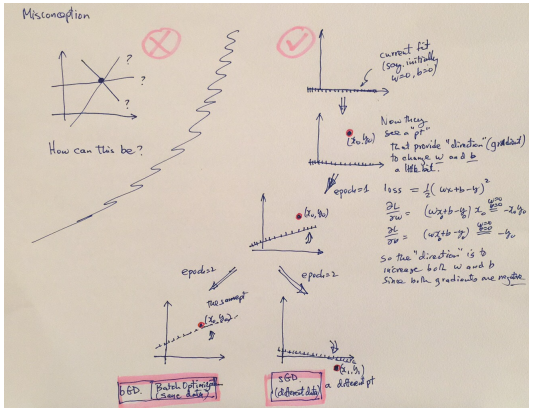


Misc
(Previous Issues)

Exercise Example 1

1 Source of Error.

Suppose that n data points $\{y_i\}$ $(i = 1, \ldots, n)$ are drawn from an independent and identical Gaussian distribution

$$p(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2}} . \quad (1)$$

Now a learning algorithm is used to estimate (or learn) the parameter μ from the data points. Work out the bias and variance of for each of the following algorithms:

(a) The "Silly" Algorithm. The algorithm ignores the data and takes a constant $\hat{\mu} = \mu_0$ as the best estimation for μ.

(b) The "Lazy" Algorithm. The algorithm takes the first data point $\hat{\mu} = y_1$ as the best estimation for μ.

(c) The "Usual" Algorithm. The standard algorithm that is consistent with our intuition (i.e., based on maximum likelihood reasoning) is to use the mean value $\hat{\mu} = \frac{1}{n}\sum_i y_i$ as the best estimation for μ.

$$< X >= \int X P(X) dX$$

$$< X >= \frac{1}{N} \sum X_i$$

$(X_1, ..., X_N)$ are the observed data generated by $P(X)$

For $\hat{\mu} = y_1$

$$\overline{\mu} = \langle y_1 \rangle_D = \int y_1 p(y_1) dy_1 = \mu$$

so Bias is $\overline{\mu} - \mu = 0$ again. Variance is computed as the following

$$Var = \langle (\hat{\mu} - \overline{\mu})^2 \rangle_D = \langle (y_1 - \mu)^2 \rangle_D = \int (y_1 - \mu)^2 p(y_1) dy_1 = 1$$

Thus this algorithm is unbiased, but not very precise because the variance is finite (Var=1).

2 Analysis of Error: Hypothesis Complexity and Sample Size.

Assume that the input variable x is uniformly distributed in the interval [-1, 1] and the target function is known $f(x) = x^2$. Of course, the target function is normally unknown to us and it is assumed to be known here so that we can carry out the analytical as well as numerical analysis on error, i.e., bias and variance.

The data set consists of 2 points $\{x_1, x_2\}$. Thus, the full data set is $D = \{(x_1, x_1^2), (x_2, x_2^2)\}$. The learning algorithm of least squares fit will return the best function, $g^D(x)$, out of the function space described by our hypothesis below that best fits the data set D.

Hypothesis 1 - Constant Functions: $\mathcal{H}_0 : h(x) = a$

Hypothesis 2 - Linear Functions: $\mathcal{H}_1 : h(x) = ax + b$

For each of the hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$,

(a) Derive the analytical expression for the average function $\overline{g}(x) = \langle g^D(x) \rangle_D$ where $\langle \ldots \rangle_D$ stands for average over data sets. In general, $g^D(x)$ is a function of x, x_1, and x_2 for a two-point data sets. Note that we don't need to state explicitly $g^D(x)$ also uses y_1 and y_2 since y_1 and y_2 themselves are functions of x_1 and x_2 respectively. Thus

$$\overline{g}(x) = \langle g^D(x) \rangle_D = \int_{x_1=-1}^{x_1=1} \int_{x_2=-1}^{x_2=1} g(x, x_1, x_2) p(x_1) p(x_2) dx_1 dx_2 \tag{2}$$

where $p(x_1) = 1/2$ and $p(x_2) = 1/2$ (i.e., uniform distribtuion over $[-1, 1]$).

(b) Calculate $Bias$, $Variance$, and MSE. Recall each term is defined as

$$Bias^2 = \langle [f(x) - \overline{g}(x)]^2 \rangle_x = \int_{x=-1}^{x=1} [f(x) - \overline{g}(x)]^2 p(x) dx \tag{3}$$

$$Var = \langle \langle [g^D(x) - \overline{g}(x)]^2 \rangle_D \rangle_x \tag{4}$$

$$MSE = \langle \langle [f(x) - g^D(x)]^2 \rangle_D \rangle_x \tag{5}$$

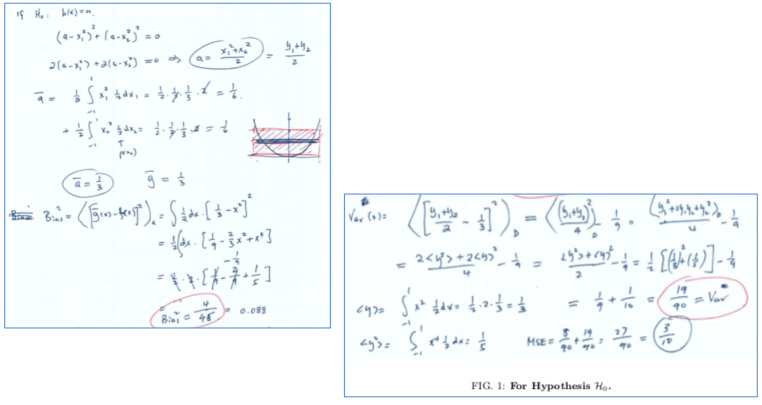


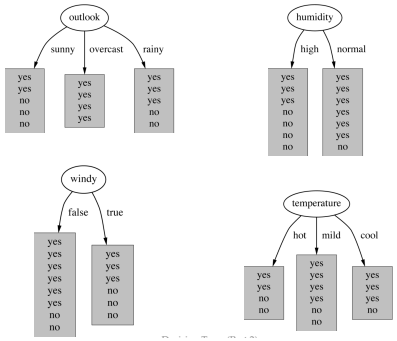
FIG. 1: **For Hypothesis $\mathcal{H}_0$.**

Decision Tree

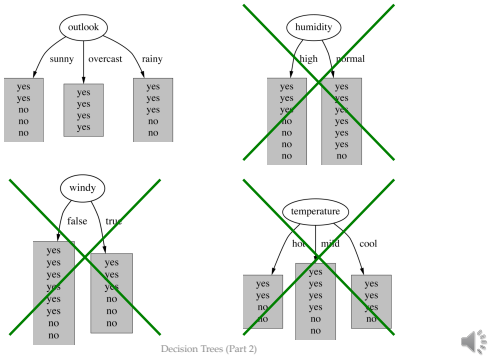
Play or not?

OUTLOOK	TEMP	HUMIDITY	WINDY	PLAY
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No%%

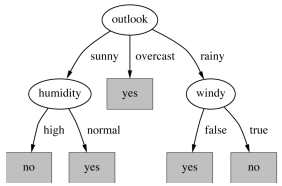
Which attribute to select?



Which attribute to select?



Final decision tree



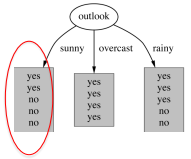
⇒ Splitting stops when data can't be split any further

Entropy: Outlook, sunny

- Formulae for computing the entropy:

$$\text{entropy}(p_1, p_2, \ldots, p_n) = -p_1 \log p_1 - p_2 \log p_2 \ldots - p_n \log p_n$$

$$H(X) = -\sum_x p(x) \log p(x)$$



$$\text{info}([2,3]) = -2/5 \times \log 2/5 - 3/5 \times \log 3/5 = 0.971 \text{ bits},$$

= (((-2) / 5) log2(2 / 5)) + (((-3) / 5) x log2(3 / 5)) = 0.97095059445

Computing Information Gain

- Information gain: information before splitting – information after splitting

  gain(*Outlook*) = info([9,5]) – info([2,3],[4,0],[3,2])
  = 0.940 – 0.693
  = 0.247 bits

- Information gain for attributes from weather data:

gain(*Outlook*)	= 0.247 bits
gain(*Temperature*)	= 0.029 bits
gain(*Humidity*)	= 0.152 bits
gain(*Windy*)	= 0.048 bits

Equation 6-1. Gini impurity

$$G_i = 1 - \sum_{k=1}^{n} p_{i,k}^2$$

Equation 6-3. Entropy

$$H_i = -\sum_{\substack{k=1 \\ p_{i,k} \neq 0}}^{n} p_{i,k} \log\left(p_{i,k}\right)$$

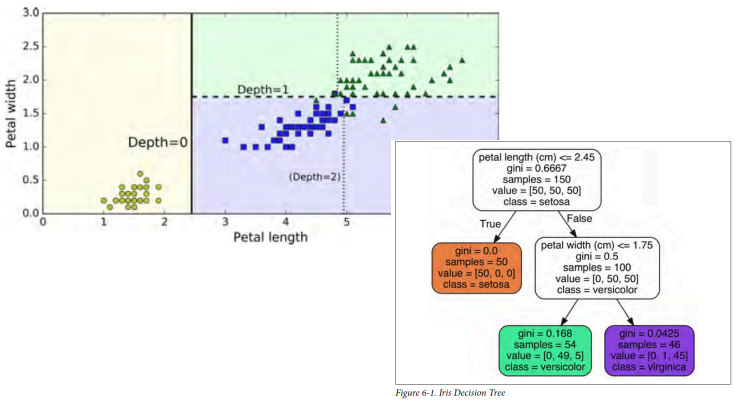
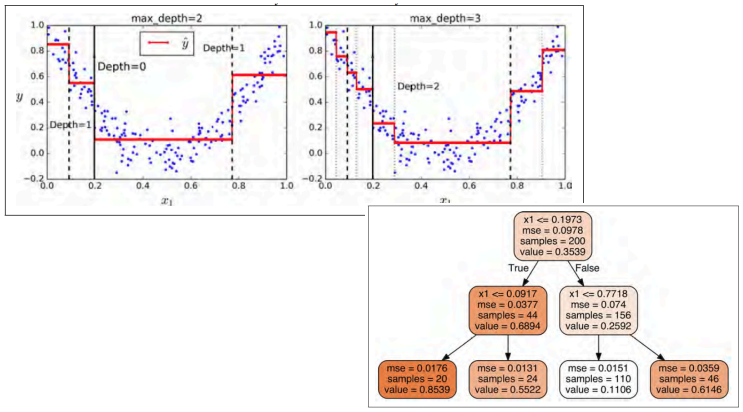


Figure 6-1. Iris Decision Tree



ID3, CART, C4.5

- Top-down induction of decision trees: ID3, algorithm developed by Ross Quinlan
  - Gain ratio just one modification of this basic algorithm
  - ⇒ C4.5: deals with numeric attributes, missing values, noisy data
- Similar approach: CART
- There are many other attribute selection criteria!

	Splitting Criteria	Attribute type	Missing values	Pruning Strategy	Outlier Detection
ID3	Information Gain	Handles only Categorical value	Do not handle missing values.	No pruning is done	Susceptible to outliers
CART	Towing Criteria	Handles both Categorical & Numeric value	Handle missing values.	Cost-Complexity pruning is used	Can handle Outliers
C4.5	Gain Ratio	Handles both Categorical & Numeric value	Handle missing values.	Error Based pruning is used	Susceptible to outliers