CPSC 633-600 Homework 3 (Total 100 points)
Decision Tree Learning

See course web page for the **due date**.
Use **elearning.tamu.edu** to submit your assignments, or submit a hard copy.

Instructor: Yoonsuck Choe

March 7, 2015

1 Entropy

Given a random variable X that can take on values $\{\oplus, \ominus\}$, the entropy is defined as:

$$E(X) = -\sum_{x \in \{\oplus, \ominus\}} P(X = x) \log_2 P(X = x).$$

Since $P(X = \oplus) + P(X = \ominus) = 1$, $E(X)$ can be rewritten as a function of $P(X = \oplus)$: Letting $p_\oplus = P(X = \oplus)$:

$$E(X) = f(p_\oplus) = -p_\oplus \log_2 p_\oplus - (1 - p_\oplus) \log_2 (1 - p_\oplus).$$

Figure 1 shows how $f(p_\oplus)$ behaves as $p_\oplus$ changes.

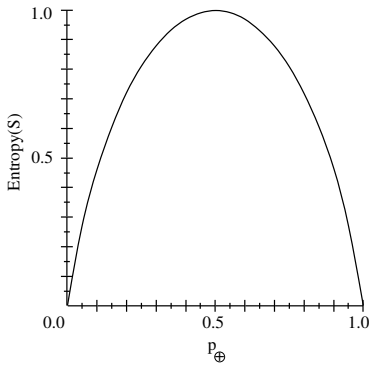

Figure 1: **Entropy.**

Problem 1 (Written: 10 pts): Extend the above analysis to a random variable Y that can take on values $\{\alpha, \beta, \gamma\}$. Given $p_\alpha = P(Y = \alpha)$, etc.,

1. Derive $E(Y)$ as a function of p_α and p_β:

   $$E(Y) = f(p_\alpha, p_\beta) = ...$$

   Note: $p_\alpha + p_\beta + p_\gamma = 1.0$.

2. For which values of p_α and p_β does $E(Y)$ become maximal (no need to derive it exactly from $f(p_\alpha, p_\beta)$–consider when it is maximal in the 2-value case)?

3. Explain why (you don't need to provide a formal proof).

Problem 2 (Program: 20 pts): Write a short program to calculate $f(p_\alpha, p_\beta)$ derived above, and obtain the $E(Y) = f(p_\alpha, p_\beta)$ values for all combinations of $p_\alpha, p_\beta \in \{0.0, 0.01, 0.02, ..., 0.99, 1.0\}$, and plot in 3D (Octave: use `surf`; Matlab: use `surf`; or draw by hand). You have to be careful because:

- log(0) will throw an error, so you have to check for the occurrence of (0 * log(0)) and make that 0 before log(0) gets evaluated. Alternatively, you can start with a value close to 0: 0.001, 0.01, 0.02, ..., 0.99, 1.0.
- Also, you have to plot for the (p_α, p_β) that sums up to less than or equal to 1.0.

Note that you can use `meshgrid` to get the grid: `[pa,pb] = meshgrid((0.001:0.01:1));`

Problem 3 (Written: 10 pts): Based on the insight from above, when you have a random variable X that can take on four different discrete values (say, $\{a, b, c, d\}$), then (1) what should $P(X = a)$ etc. be so that the entropy of X is maximized? (2) Also, what is the value of the maximum entropy? (Note: You don't need to provide a formal proof.)

2 Decision Tree Learning (ID3)

Problem 4 (Written: 30 pts): Calculate the following **by hand** and show all intermediate results (you may use a calculator for intermediate numerical results). (1) Calculate the entropy of the following training set. (2) Calculate the information gain for each of the three attributes. (3) Which one is the best attribute to test first?

Instance#	Age	Sex	Breed	Decision (Adopt?)
1	2 years old	Male	Pomeranian	N
2	1 year old	Male	Chihuahua	Y
3	4 years old	Female	Australian Shepherd	Y
4	2 years old	Male	Pit Bull	N
5	1 year old	Male	Australian Shepherd	Y
6	1 year old	Male	Pit Bull	N
7	1 year old	Female	Australian Shepherd	N
8	1 year old	Female	Chihuahua	Y
9	4 years old	Female	Pomeranian	N
10	2 years old	Male	Chihuahua	Y
11	2 years old	Female	Pomeranian	Y
12	2 years old	Female	Australian Shepherd	N

Problem 5 (Program: 30 pts): Write a program for decision tree learning algorithm and apply it to the simple data set above and show the resulting decision tree. You may use any programming language for the learning part.

Use graphviz (`http://www.graphviz.org`, available on all platforms) to plot the decision tree. Generate code for the graph within your decision tree learning program, and run graphviz's tool called `dot` to visualize. The syntax for `dot` is very simple, and you only need to use the basic features (`http://www.graphviz.org/content/dot-language`).

```
digraph G {

  // attributes
  attr1 [shape="rectangle", label="Test"]
  attr2 [shape="rectangle", label="Exam"]
  attr3 [shape="rectangle", label="Check"]

  // leaves
  leaf1 [shape="plaintext", label="Yes"]
  leaf2 [shape="plaintext", label="No"]
  leaf3 [shape="plaintext", label="No"]
  leaf4 [shape="plaintext", label="Yes"]
  leaf5 [shape="plaintext", label="Yes"]

  // connections
  attr1 -> attr2 [label="Tval 1"]
  attr1 -> leaf1 [label="Tval 2"]
  attr1 -> leaf5 [label="Tval 3"]
  attr2 -> leaf2 [label="Eval a"]
  attr2 -> attr3 [label="Eval b"]
  attr3 -> leaf3 [label="Cval I"]
  attr3 -> leaf4 [label="Cval II"]

}
```

Save the above to a file `dt.dot` and run `dot -Tpng -o dt.png dt.dot`, and check the resulting file `dt.png`.