CS 189 Introduction to Machine Learning
Spring 2020 Jonathan Shewchuk

DIS9

1 Surprise and Entropy

In this section, we will clarify the concepts of surprise and entropy. Recall that entropy is one of the standards for us to split the nodes in decision trees until we reach a certain level of homogeneity.

(a) Suppose you have a bag of balls, all of which are black. How surprised are you if you take out a black ball?

Solution: 0. We aren't surprised at all when events with probability 1 occur.

(b) With the same bag of balls, how surprised are you if you take out a white ball?

Solution: ∞. We are infinitely surprised when an event with probability 0 occurs.

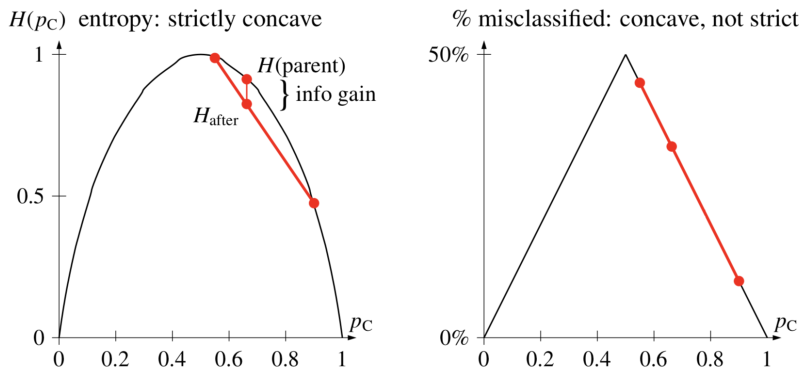
(c) Now we have 10 balls in the bag, each of which is black or white. Under what color distribution(s) is the entropy of the bag minimized? And under what color distribution(s) is the entropy maximized? Calculate the entropy in each case.
Recall: The entropy of an index set S is a measure of expected surprise from choosing an element from S; that is,

$$H(S) = -\sum_C p_C \log_2(p_C), \text{ where } p_C = \frac{|i \in S : y_i = C|}{|S|}.$$

Solution: The entropy is minimized when, for example, all the balls are black or all the balls are white. In this case the entropy is 0. The entropy is maximized when half the balls are black and half the balls are white, in which case the entropy is $-0.5\log_2 0.5 - 0.5\log_2 0.5 = 1$.

(d) Draw the graph of entropy $H(p_c)$ when there are only two classes C and D, with $p_D = 1 - p_C$. Is the entropy function strictly concave, concave, strictly convex, or convex? Why? What is the significance?
Hint: For the significance, recall the information gain.

Solution: The function is strictly concave. Notice that the function $-x\log x$ is strictly concave in $[0, 1]$, and a sum of strictly concave functions is strictly concave.
Significance: (from lecture) Suppose we pick two points on the entropy curve, then draw a line segment connecting them. Because the entropy curve is strictly concave, the interior of the line segment is strictly below the curve. Any point on that segment represents a weighted average of the two entropies for suitable weights. If you unite the two sets into one parent set, the parent set's value p_C is the weighted average of the children's p_c's. Therefore, the point



directly above that point on the curve represents the parent's entropy. The information gain is the vertical distance between them. So the information gain is positive unless the two child sets both have exactly the same p_C and lie at the same point on the curve.

On the other hand, for the graph on the right, plotting the % misclassified, if we draw a line segment connecting two points on the curve, the segment might lie entirely on the curve. In that case, uniting the two child sets into one, or splitting the parent set into two, changes neither the total misclassified sample points nor the weighted average of the % misclassified. The bigger problem, though, is that many different splits will get the same weighted average cost; this test doesn't distinguish the quality of different splits well.

2 Decision Trees and Random Forests

Random forests are a specific ensemble method where the individual models are decision trees trained in a randomized way so as to reduce correlation among them. Because the basic decision tree building algorithm is deterministic, it produces the same tree every time if we give it the same dataset and use the same hyperparameters (stopping conditions, etc.).

Consider constructing a multi-class binary decision tree on n training points with d real-valued features. The splits are chosen to maximize the information gain. We only consider splits that form a linear boundary orthogonal to one of the coordinate axes.

(a) Give an example or disprove: For every $n \geq 3$, there exists some discrete probability distribution on n objects whose entropy is negative.

Solution: False. The entropy is always nonnegative since $-p \log_2 p$ is nonnegative when $p \in (0, 1]$. (For an object with probability $p = 0$, we can either ignore the object or adopt the convention that $0 \log 0 = 0$, as $\lim_{n \to 0^+} = 0$.)

(b) One may be concerned that the randomness introduced in random forests may cause trouble. For example, some features or sample points may never be considered at all. We investigate this phenomenon in parts (b)–(d). Consider n training points in a feature space of d dimensions. Consider building a random forest with T binary trees, each having exactly h internal nodes. Let m be the number of features randomly selected (from among d input features) at each treenode. For this setting, compute the probability that a certain feature (say, the first feature) is never considered for splitting in any treenode in the forest.

Solution: The probability that it is not considered for splitting in a particular node of a particular tree is $1 - \frac{m}{d}$. The subsampling of m features at each treenode is independent of all others. There is a total of ht treenodes and hence the final answer is $(1 - \frac{m}{d})^{hT}$.

(c) Now let us investigate the possibility that some sample point might never be selected. Suppose each tree employs $n' = n$ bootstrapped (sampled with replacement) training sample points. Compute the probability that a particular sample point (say, the first sample point) is never considered in any of the trees.

Solution: The probability that it is not considered in one of the trees is $(1 - \frac{1}{n})^n$, which approaches $1/e$ as $n \to \infty$. Since the choice for every tree is independent, the probability that it is not considered in any of the trees is $(1 - \frac{1}{n})^{nT}$, which approaches e^{-T} as $n \to \infty$.

(d) Compute the values of the two probabilities you obtained in parts (b) and (c) for the case where there are $n = 2$ training points with $d = 2$ features each, $T = 10$ trees with $h = 4$ internal nodes each, and we randomly select $m = 1$ potential splitting features in each treenode. You may leave your answer in a fraction and exponentiated form, e.g., $\left(\frac{51}{100}\right)^2$. What conclusions can you draw about the concern mentioned in part (b)?

Solution: $\frac{1}{2^{40}}$ and $\frac{1}{2^{20}}$. It is quite unlikely that a feature or a sample will be missed.

3 Decision Boundary Visualization on Decision Tree and Random Forest

In this problem, we will visualize the decision boundaries of decision tree, random forest, and adaboost with decision tree. Please go to the Jupyter Notebook part and visualize the decision boundaries of the above approaches. You do not need to write code in the Jupyter Notebook. (Use this notebook to open the file in the Google drive and follow instructions therein).