CS 194-10, Fall 2011
Assignment 3 Solutions

1. *Entropy and Information Gain*

   (a) To prove $H(S) \leq 1$, we can find the global maximum of $B(S)$ and show that it is at most 1. Since $B(q)$ is differentiable, we can set the derivative to 0,

   $$0 = \frac{\partial B}{\partial q} = -\log q - 1 + \log(1-q) + 1$$

   which yields $q = 0.5$. Noting that entropy is concave, we get a global maximum by plugging this value. Therefore $H(q) \leq 1$, and we have equality when $q = p/(p+n) = 1, i.e., p = n$.

   (b) This result emphasizes the fact that any statistical fluctuations caused by the random sampling process will result in an apparent information gain.

   The easy part is showing that the gain is zero when each subset has the same ratio of positive examples. Since $p = \sum p_k$ and $n = \sum n_k$, if $p_k/(p_k + n_k)$ is the same for all k we must have $p_k/(p_k + n_k) = p/(p+n)$ for all k. From this, we obtain

   $$\begin{aligned} Gain &= B\left(\frac{p}{p+n}\right) - B\left(\frac{p}{p+n}\right)\frac{1}{p+n}\sum_{k=1}^{d} p_k + n_k \\ &= B\left(\frac{p}{p+n}\right) - B\left(\frac{p}{p+n}\right)\frac{1}{p+n}(p+n) = 0 \end{aligned}$$

   Note that this holds for all values of $p_k + n_k$. To prove that the value is positive elsewhere, we can apply the method of Lagrange multipliers to show that this is the only stationary point; the gain is clearly positive at the extreme values, so it is positive everywhere but the stationary point. In detail, we have constraints $\sum_k p_k = p$ and $\sum_k n_k = n$, and the Lagrange function is

   $$\Lambda = B\left(\frac{p}{p+n}\right) - \sum_k \frac{p_k + n_k}{p+n} B\left(\frac{p_k}{p_k + n_k}\right) + \lambda_1\left(p - \sum_k p_k\right) + \lambda_2\left(n - \sum_k n_k\right) .$$

   Setting its derivatives to zero, we obtain, for each k,

   $$\begin{aligned} \frac{\partial \Lambda}{\partial p_k} &= -\frac{1}{p+n}B\left(\frac{p_k}{p_k+n_k}\right) - \frac{p_k+n_k}{p+n}\log\frac{p_k}{n_k}\left(\frac{1}{p_k+n_k} - \frac{p_k}{(p_k+n_k)^2}\right) - \lambda_1 = 0 \\ \frac{\partial \Lambda}{\partial n_k} &= -\frac{1}{p+n}B\left(\frac{p_k}{p_k+n_k}\right) - \frac{p_k+n_k}{p+n}\log\frac{p_k}{n_k}\left(\frac{-p_k}{(p_k+n_k)^2}\right) - \lambda_2 = 0 . \end{aligned}$$

   Subtracting these two, we obtain $\log(p_k/n_k) = (p+n)(\lambda_2 - \lambda_1)$ for all k, implying that at any stationary point the ratios p_k/n_k must be the same for all k. Given the two summation constraints, the only solution is the one given in the question.

2. *Empirical Loss and Splits*

   (a) 0/1 Loss. Let p_k and n_k be the number of positive and negative examples respectively for each subset. Then the loss for the parent is $\min(\sum_k p_k, \sum_k n_k)$ and the total loss for the children is given by $\sum_k \min(p_k, n_k)$. We'd like to show that,

   $$\sum_k \min(p_k, n_k) \leq \min(\sum_k p_k, \sum_k n_k) \quad . \tag{1}$$

Note that $\min(p_k, n_k) \leq p_k$, therefore $\sum_k \min(p_k, n_k) \leq \sum_k p_k$. Similarly $\sum_k \min(p_k, n_k) \leq \sum_k n_k$. Hence the assertion in (1) is correct, i.e., 0/1 loss can never increase when splitting.

(b) L_2 loss is minimzed in any given set by returning the sample mean $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$, giving L_2 loss $\sum_{i=1}^{N}(y_i - \bar{y})^2$. Suppose we split the set into subsets A and B, with sample means $\bar{y}_A$ and $\bar{y}_B$. Since each minimizes the L_2 loss for its respective subset, we have

$$\sum_{j \in A}(y_j - \bar{y}_A)^2 \leq \sum_{j \in A}(y_j - \bar{y})^2 \quad \text{and} \quad \sum_{k \in B}(y_k - \bar{y}_B)^2 \leq \sum_{k \in B}(y_k - \bar{y})^2 \ .$$

Adding these two inequalities, we obtain

$$\sum_{j \in A}(y_j - \bar{y}_A)^2 + \sum_{k \in B}(y_k - \bar{y}_B)^2 \leq \sum_{j \in A}(y_j - \bar{y})^2 + \sum_{k \in B}(y_k - \bar{y})^2 = \sum_{i=1}^{N}(y_i - \bar{y})^2$$

hence the L_2 loss cannot increase.

3. *Splitting continuous attributes*
Without loss of generality, consider a node with p positive and n negative examples and $p \leq n$. Let p_k and n_k be the number of positive and negative examples after a split. Consider a case where we split between positive examples such that the child node on the left is more positive and the child node on the right is more negative, i.e. $p_k \geq n_k$ and $p_{k+1} \leq n_{k+1}$. The empirical loss for the two child nodes is $n_k + p_{k+1}$. We can improve the empirical loss by moving the split one position to the right, which effectively takes a wrongly classified example from the right child node and turns it into a correctly classified example in the left node. Majority is still maintained and the empirical loss becomes $n_k + p_{k+1} - 1$. We can repeatedly apply this argument to see that the optimal split must occur at the dividing point between samples of different classes.

4. *Majority voting*

(a) In order for the majority vote classifier to make a mistake, more than half of the K classifiers must fail. Since each classifier fails independently with Bernoulli(ϵ), the probability that more than $K/2$ out of K trials of independent Bernoulli(ϵ) variables are 1 gives the desired probability:

$$\epsilon_{majority} = \sum_{n=\lfloor K/2 \rfloor + 1}^{K} \binom{K}{n} \epsilon^n (1-\epsilon)^{K-n}$$

(b) Yes, if the independence assumption is removed, the ensemble error can be worse than ϵ. Consider the case where we have 3 classifiers A,B,C and consider the following scenario where ✓ is a correct prediction and × is an incorrect prediction.

A	B	C	$Majority$
×	×	✓	×
✓	×	×	×
×	✓	×	×
✓	✓	✓	✓
...	...	...	...

(2)

If the above pattern continues forever, we get $\epsilon_{majority} = 3/4$ while $\epsilon = 1/2$.