STOR 565 Homework

1. Show that the following functions $f, g, h : [0,1] \to \mathbb{R}$ used to define impurity measures for growing trees are concave.

   a. $m(p) = \min(p, 1-p)$

   b. $g(p) = p(1-p)$

   c. $h(p) = -p \log p - (1-p)\log(1-p)$, with the convention that $0 \log 0 = 0$

Which of these functions is strictly concave?

2. Let $D_n = (x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \{0,1\}$ be a data set for classification. For each region $A \subseteq \mathcal{X}$ let $|A|$ denote the number of points x_i in A and let $p(A) = |A|^{-1} \sum_{x_i \in A} y_i$ be the fraction of points $x_i \in A$ labeled 1. Suppose that the region A can be expressed as the disjoint union $A = A_1 \cup A_2$ of two other regions.

   a. Show that

   $$p(A) \;=\; \frac{|A_1|}{|A|} p(A_1) + \frac{|A_2|}{|A|} p(A_2)$$

   b. Conclude from (a) that for any concave function $f : [0,1] \to \mathbb{R}$

   $$f(p(A)) - \frac{|A_1|}{|A|} f(p(A_1)) - \frac{|A_2|}{|A|} f(p(A_2)) \;\geq\; 0$$

   This establishes that the impurity differences defined in the lecture for the misclassification, Gini, and entropy impurity measures are non-negative.

   c. Let $m(p) = \min(p, 1-p)$. Show that $|A| m(p(A))$ is the number of misclassifications if every point in A is assigned to the majority class.

   d. Consider two partitions γ_1 and γ_2 of $\mathcal{X}$ that are identical except that a cell A of γ_1 is split into two cells A_1 and A_2 in γ_2. What can you say about the training error of the corresponding histogram classification rules (based on majority voting in cells)?

3. Let $D_n = (x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \{0,1\}$ be a data set for classification and let $\gamma = \{A_1, \ldots, A_K\}$ be a partition of $\mathcal{X}$. Define the histogram classification rule $\hat{\phi}_\gamma$ based on γ. Show that $\hat{\phi}_\gamma$ minimizes the training error $R_n(\phi)$ over all classification rules ϕ that are constant on the cells of γ, meaning $\phi(u) = \phi(v)$ if u, v are in the same cell of γ.