Problem 1: No Free Lunch Theorem (25%)

This is problem 1 from chapter 9 DHS.

One of the “conservation laws” for generalization states that the positive generalization performance of an algorithm in some learning situations must be offset by the negative performance elsewhere. Consider a very simple learning algorithm that seems to contradict this law. For each test pattern, the prediction of the majority learning algorithm is merely the category most prevalent in the training data.

1. Show that averaged over all two-category problems of a given number of features, the off-training set error is 0.5.
2. Repeat (1) by for the minority learning algorithm, which always predicts the label of the category least prevalent in the training data.
3. Use your answers from (1) and (2) to illustrate part 2 of the No Free Lunch Theorem.

Problem 2: Jackknife Statistics (25%)

These are problems 23 and 26 from chapter 9 DHS.

Show that the jackknife estimates of the mean and the variance,

\[ \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \mu_{(i)} \]

\[ \text{Var}[\hat{\mu}] = \frac{(n-1)}{n} \sum_{i=1}^{n} (\mu_{(i)} - \mu_{(\cdot)}) \]

respectively, where each \( \mu_{(i)} \) is a leave-one-out mean,

\[ \mu_{(i)} = \frac{1}{n-1} \sum_{j \neq i} x_j \]

are equivalent to the traditional estimates of the mean and the variance (sample mean, unbiased sample variance).

Problem 3: Kernel Density Estimation (25%)

This is problem 2 from chapter 4 DHS.

Consider a normal \( p(x) \sim N(\mu, \sigma^2) \) and Parzen window function \( \varphi(x) \sim N(0, 1) \). Show that the Parzen window estimate

\[ p_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} \varphi \left( \frac{x - x_i}{h_n} \right) \]

has the following properties:
1. \( \bar{p}_n(x) \sim N(\mu, \sigma^2 + h_n^2) \)

2. \( \text{var}[p_n(x)] \simeq \frac{1}{2n h_n \sqrt{\pi}} p(x) \)

3. \( p(x) - \bar{p}_n(x) \simeq \frac{1}{2} \left( \frac{h_n}{\sigma} \right)^2 \left[ 1 - \left( \frac{x - \mu}{\sigma} \right)^2 \right] p(x) \)

for small \( h_n \). (Note that if \( h_n = h_1 / \sqrt{n} \), this result implies that the error due to bias goes to zero as \( 1/n \), whereas the standard deviation of the noise only goes to zero as \( \sqrt{n} \).

**Problem 4: Ensemble Training Error in AdaBoost (25%)**

*This is part of problem 32 from chapter 9 DHS.*

Consider AdaBoost with an arbitrary number of weak classifiers.

1. State clearly any assumptions you make, and derive

\[
E = \prod_{k=1}^{k_{\text{max}}} \left[ 2 \sqrt{E_k (1 - E_k)} \right] = \prod_{k=1}^{k_{\text{max}}} \sqrt{1 - 4G_k^2} \leq \exp \left( -2 \sum_{k=1}^{k_{\text{max}}} G_k^2 \right)
\]

for the ensemble training error of the full boosted system.

2. Recall that the training error for a weak learner applied to a two-category problem can be written \( E_k = \frac{1}{2} - G_k \) for some positive value \( G_k \). The training error for the first component classifier is \( E_1 = 0.25 \). Suppose that \( G_k = 0.05 \) for all \( k = 1 \) to \( k_{\text{max}} \). Plot the upper bound on the ensemble test error given by the equation above (which you just derived), such as the plot shown in DHS Fig. 9.7.