1. Assume you have a dataset $S$ consisting of $n$ instantiations of the random variable $X \sim P$. Every example $x_i \in S$ has exactly one attribute (it’s scalar). We estimate the mean as $E[X] \approx \frac{1}{n} \sum_{i=1}^{n} x_i = m$ and the variance as $V(X) = E[X^2] - E[X]^2 \approx \frac{1}{n} \sum_{i=1}^{n} x_i^2 - m^2 = \nu$.

Next we define a new random variable $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$, where every $X_i$ is distributed as $P$. We will be interested in the distribution $Q$ for $\overline{X}$. Due to the central limit theorem we expect $Q$ to be approximately Gaussian for large enough $n$.

a) Show that $E[\overline{X}] = E[X] \approx m$

b) We define the variance $V[\overline{X}] = E[\overline{X}^2] - E[\overline{X}]^2$.

Show that $V[\overline{X}] = \frac{1}{n} \left( E[X^2] - E[X]^2 \right) \approx \frac{\nu}{n}$

2. Last week you had implemented 1NN and LR for the Iris data.

a) For both algorithms compute the 95% confidence intervals for the classification error (sometimes called “standard error”). Use 10 (90%-10%) splits of the data and average your results over these 10 splits.

b) Now implement the algorithm explained in class to compare these two algorithms and see if you can conclude with 95% confidence that one algorithm performs superior to another algorithm (this involves doing a paired t-test on the 10 splits of the data). You may use the mfile “ttest.m” in matlab to perform the paired t-test.

3. Call $h(x,D)$ a classifier learned on dataset $D$, and which we test on query point $x$. Call $f(x)$ the ground truth classifier which we try to estimate. We will be interested in the total squared error of this classifier:

$$SE(x) = E[(h(x,D) - f(x))^2]_{P(D)}$$

where we take expectation over infinitely many datasets $D$, drawn from $P(D)$.
a) Show that the following bias-variance decomposition can be derived:

\[ SE(x) = E((h(x, D) - E[h(x, D)])^2) + (f(x) - E[h(x, D)])^2 \]

b) Which terms represent the bias and which term the variance?

c) If we have a very flexible hypothesis class \( H \) from which we try to learn \( h(x) \), which of the two terms will dominate our squared error?