Statistics for Evaluating the Performance of Machine Learning Models

We can compare two supervised machine learning models using the same training dataset by applying the classical hypothesis testing paradigm. Our measure of model performance is the test-set error rate. Below, we state the problem in the form of a *null hypothesis*.

> There is no significant difference in the test set error rate of two supervised machine learning models, M_1 and M_2, built with the same training dataset.

Although there are several possible test set scenarios we can use, we will take the approach that allows us to compare performance between models when:

- *the **same** test dataset is used to compare the models, and*
- *the comparison is is based on the overall classsification correctness of the models.*

While this is a simple approach, it would perhaps be more straight forward to compare model performance using **two independent testing datasets selected randomly from a a pool of sample data.** However, this approach is feasible only if there is a large supply of data available from which random test sets can be extracted. Thus, when smaller-sized datasets are all we have to work with, a single test set used on both models is probably the only possibility.

Comparing Models with Independent Test Data

The simpler approach that we will take is to use a technique that compares the overall classification correctness of the models. *It shoud be noted that this method is equally valid wheter we have two independent test datasets or only a single dataset to use with both models.*

The most general form of the statistic for comparing the performnce of two classifier modesl M_1 and M_2 is

$$P = \frac{|E_1 - E_2|}{\sqrt{q(1-q)(1/n_1 + 1/n_2)}}$$

where:

E_1	is the error rate for model M_1
E_2	is the error rate for model M_2
q	$= \frac{(E_1+E_2)}{2}$ is the error rate average
n_1	is the number of instances in test set A
n_2	is the number of instances in test set B

You should note that $q(1-q)$ is the variance score computed using the **average** of the two error rates. When we have a *single* test set of size n that we use for both models, then the formula simplifies to:

$$P = \frac{|E_1 - E_2|}{\sqrt{q(1-q)(2/n)}}$$

With either of the above equations, if, for the value of P we find $P \geq 2$, then we can be 95% confident that the difference in the test set performance is ststistically significant.

In this project, we shall use this simplified version of the relation since we only have one test dataset for both models.

BONUS (15pts): Comparing the Performance of two Supervised Classification Models

Model Performance Comparision of two Classifiers

Using the **logistic regression model** and the **KNN model** (do NOT re-instantiate and re-train the model. Use the last knn model from your previous section), compare the performasnce of the models using the above statistic when we only have one test set. To to this you need to calculate (**and print**) in individual notebook cells the following:

E_1 : the error rate for model M_1, your K-Nearest Neighbor Model

E_2 is the error rate for model M_2, your Logistic Regression Model

$q = \frac{(E_1+E_2)}{2}$, the error rate average

n_1 , the number of instances in test set A

n_2 , the number of instances in test set B

Using Python, write a simple decision construct to compute and display P, depending on whether $P \geq 2$. Use a complete sentence in your output.
