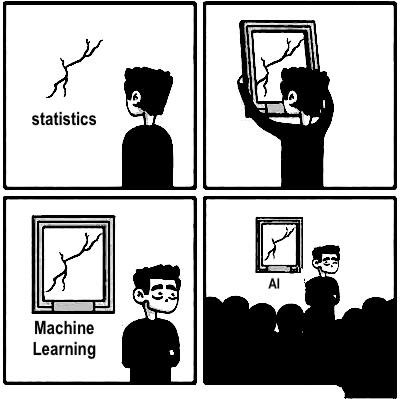
**MACHINE LEARNING WORKSHEET**



1. **Flexible vs Inflexible Methods**

For each of parts (a) through (d), indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.

1. The sample size n is extremely large, and the number of predictors p is small.
2. The number of predictors p is extremely large, and the number of observations n is small.
3. The relationship between the predictors and response is highly non-linear.
4. The variance of the error terms, i.e. Var(ϵ), is extremely high.

2. **Scenarios - Classification or Regression?**

Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

1. We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.
2. We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.
3. We are interested in predicting the % change in the US Dollar in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.

3. **Bias-Variance Decomposition**

1. Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, and the y-axis should represent the values for each curve. There should be five curves. Make sure to label each one.
2. Explain why each of the five curves has the shape displayed in part (a).

4. **Applications of Statistical Learning**

1. Describe three real-life applications in which classiﬁcation might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.
2. Describe three real-life applications in which regression might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.
3. Describe three real-life applications in which cluster analysis might be useful.

5. **Flexible vs Inflexible Methods**

What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under what circumstances might a more ﬂexible approach be preferred to a less ﬂexible approach? When might a less ﬂexible approach be preferred?

6. **Parametric & Non-Parametric Approaches**

Describe the diﬀerences between a parametric and a non-parametric statistical learning approach. What are the advantages of a parametric approach to regression or classiﬁcation (as opposed to a nonparametric approach)? What are its disadvantages?

7. **APPLIED: The College Dataset**

This exercise relates to the **College** data set, which can be found in the file [College.csv](http://faculty.marshall.usc.edu/gareth-james/ISL/data.html). It contains a number of variables for 777 different universities and colleges in the US.

The variables are:

* Private : Public/private indicator
* Apps : Number of applications received
* Accept : Number of applicants accepted
* Enroll : Number of new students enrolled
* Top10perc : New students from top 10% of high school class
* Top25perc : New students from top 25% of high school class
* F.Undergrad : Number of full-time undergraduates
* P.Undergrad : Number of part-time undergraduates
* Outstate : Out-of-state tuition
* Room.Board : Room and board costs
* Books : Estimated book costs
* Personal : Estimated personal spending
* PhD : Percent of faculty with Ph.D.’s
* Terminal : Percent of faculty with terminal degree
* S.F.Ratio : Student/faculty ratio
* perc.alumni : Percent of alumni who donate
* Expend : Instructional expenditure per student
* Grad.Rate : Graduation rate

(a) **Import**

Use the read.csv() function to read the data into R. Call the loaded data college. Make sure that you have the directory set to the correct location for the data.

(b) **View the row names**.

(c) **Exploring**

1. Use the summary() function to produce a numerical summary of the variables in the data set.
2. Use the pairs() function to produce a scatterplot matrix of the first ten columns or variables of the data. Recall that you can reference the ﬁrst ten columns of a matrix A using A[,1:10].
3. Use the plot() function to produce side-by-side boxplots of Outstate versus Private.
4. Create a new qualitative variable, called Elite, by *binning* the Top10perc variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 10% of their high school classes exceeds 50%.
5. Use the hist() function to produce some histograms with differing numbers of bins for a few of the quantitative variables.
6. Continue exploring the data, and provide a brief summary of what you discover.

8. **APPLIED: The College Dataset**

1. This exercise relates to the **College** data set, which can be found in the file [College.csv](http://faculty.marshall.usc.edu/gareth-james/ISL/data.html). It contains a number of variables for 777 different universities and colleges in the US.