INFO 7390 –
Advances in Data Sciences and Architecture
Practice Exam Solutions

Student Name: ____
Professor: Nik Bear Brown

Rules:
1. NO COMPUTER, NO PHONE, NO DISCUSSION or SHARING.
2. Ask if you don't understand a question.
3. You may have five 8½"×11" sheets of notes (you may use both sides, written or printed as small as you like).
4. Time allowed. 1 hour 30 minutes.
5. Bring pen/pencil. The midterm will be written on paper.

Q1 (5 Points) For a MLP, the number of nodes in the input layer is 10 and the hidden layer is 5. What is the maximum number of connections from the input layer to the hidden layer?

Solution:

50

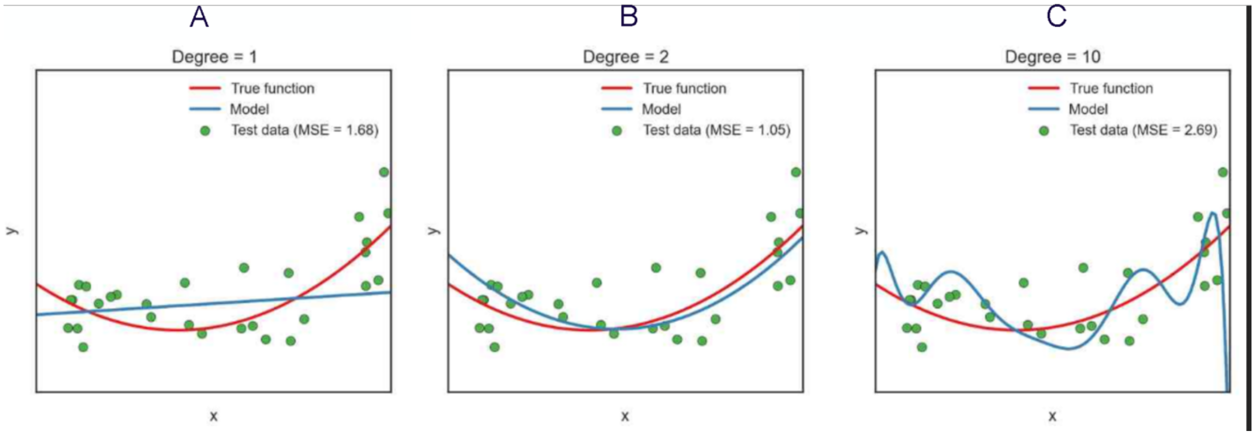
Since MLP is a fully connected directed graph, the number of connections are a multiple of number of nodes in input layer and hidden layer.

Q2 (5 Points) Does a linear regression model have less bias than a quadratic regression model?

Solution:

The linear regression model will always have more bias than a quadratic regression model as a second order polynomial will fit at least as well as a first order polynomial as since we can always set the second order term to 0.

Q3 (5 Points) The models A, B and C were fit to the same data. Are any of the three models below underfitting or overfitting?



Solution:

Model A is clearly underfitting as one sees the model is too simple to fit the test data as well as model B. Model C is clearly overfitting as one sees the poor fit on test data as compared to model B. Model B is unknown as one would need to test a slightly more complex, say degree 3, model to know if it is underfitting.

Q4 (5 Points) A feature F1 can take certain value: A, B, C, D, E, & F and represents grade of students from a college.

1) Which of the following statement is true in following case?

A) Feature F1 is an example of nominal variable.
B) Feature F1 is an example of ordinal variable.
C) It doesn't belong to any of the above category.
D) Both of these

Solution:

Solution: (B)

Ordinal variables are the variables which has some order in their categories. For example, grade A is a higher grade than grade B.

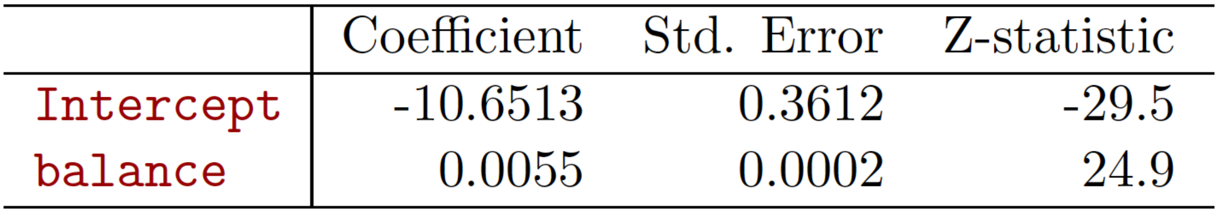
Q5 (5 Points) How does regression differ from classification?

Solution:

Regression: A supervised problem is said to be regression problem when the output variable is a numeric value such as "weight", "height" or "dollars." It can also be a probability.

Classification: It is said to be a classification problem when the output variable is a discrete (or category) such as "male" and "female" or "disease" and "no disease."

Q6 (5 Points) Assume logistic regression is being used to predict whether someone will default on a loan and an independent variable called balance is the only independent variable in the model. It returns the stats below. Is balance a significant predictor of default?



	Coefficient	Std. Error	Z-statistic
Intercept	-10.6513	0.3612	-29.5
balance	0.0055	0.0002	24.9

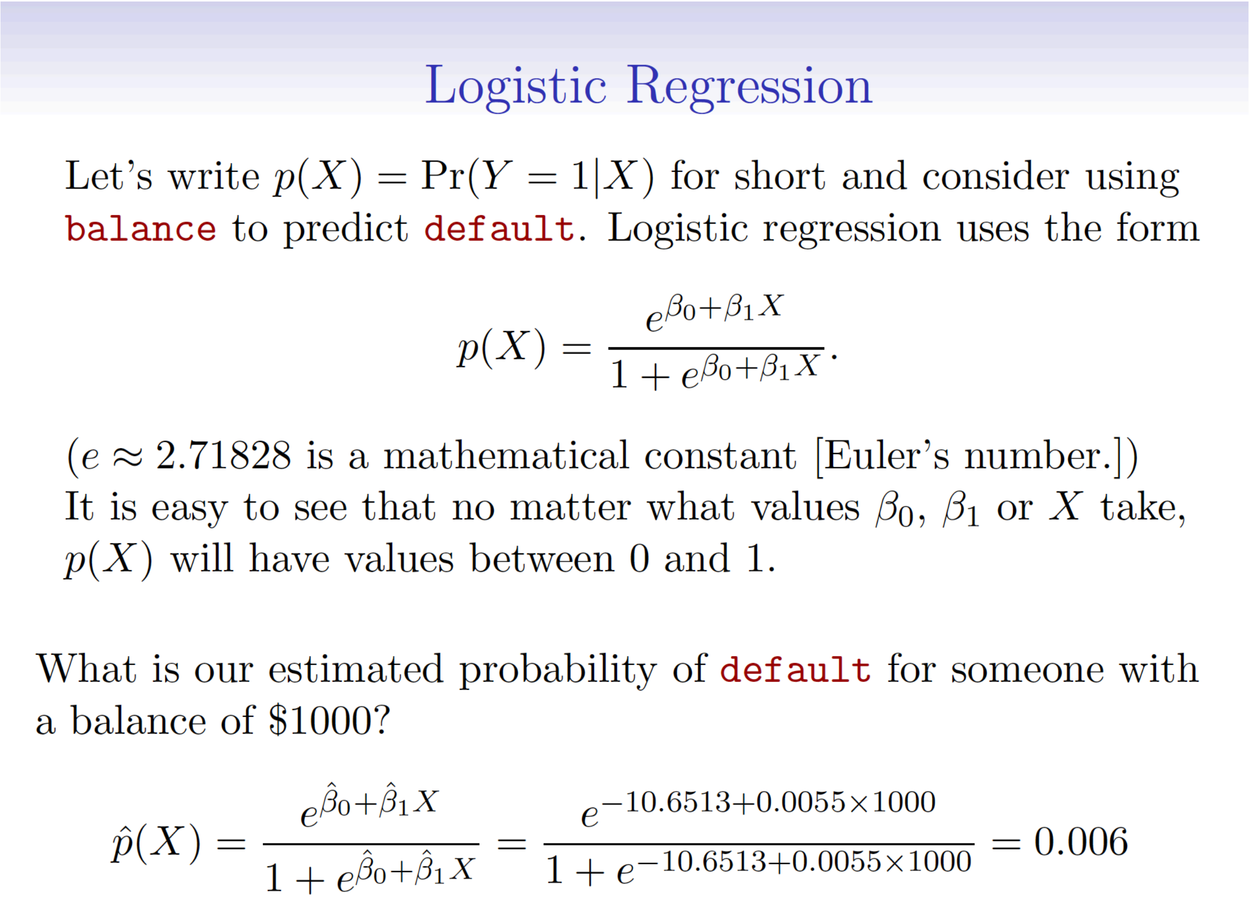
Solution:

Yes. The z-score of 24.9 for balance is very significant, about 25 standard deviations from the mean.

Q7 (5 Points) Calculate the probability of defaulting given a balance of 1000 from the stats in Q6.

Solution:

Just plug the intercept beta zero and slope beta one in to the equation below; along with an X of 1000.



Logistic Regression

Let's write $p(X) = \Pr(Y = 1|X)$ for short and consider using `balance` to predict `default`. Logistic regression uses the form

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}.$$

($e \approx 2.71828$ is a mathematical constant [Euler's number.])
It is easy to see that no matter what values β_0, β_1 or X take, $p(X)$ will have values between 0 and 1.

What is our estimated probability of `default` for someone with a balance of $1000?

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1000}}{1+e^{-10.6513+0.0055\times 1000}} = 0.006$$

Q8 (5 Points)

Explain whether each scenario is a classification or regression problem. Indicate whether we are most interested in inference or prediction.

(a) 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.

(b) 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.

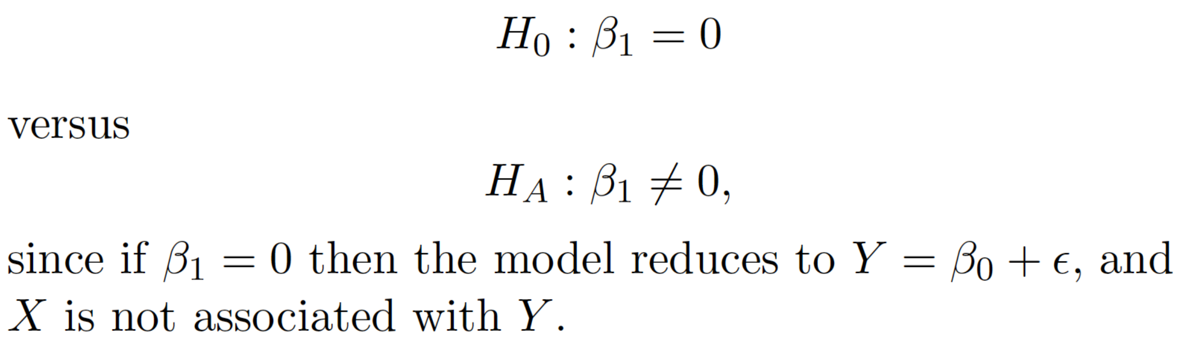
Solution:

(a) Inference because we are interested in understanding which factors
affect CEO salary. Regression because our dependent variable CEO salary is numeric.

(b) Prediction because we wish to know whether it will be a success or a failure. Classification if we predict success or a failure. Logistic regression if we predict a probability of success or a failure.

Q9 (5 Points) Describe the null hypotheses to which the p-values given in linear regression correspond.

Solution:



$$H_0 : \beta_1 = 0$$

versus

$$H_A : \beta_1 \neq 0,$$

since if $\beta_1 = 0$ then the model reduces to $Y = \beta_0 + \epsilon$, and X is not associated with Y.

Q10 (5 Points) What are:

- True positive rate (TPR),
- False positive rate (FPR),

How do TPR and FPR relate to AUC or AUROC?

Solution:

AUC = Area Under the Curve.
AUROC = Area Under the Receiver Operating Characteristic curve.

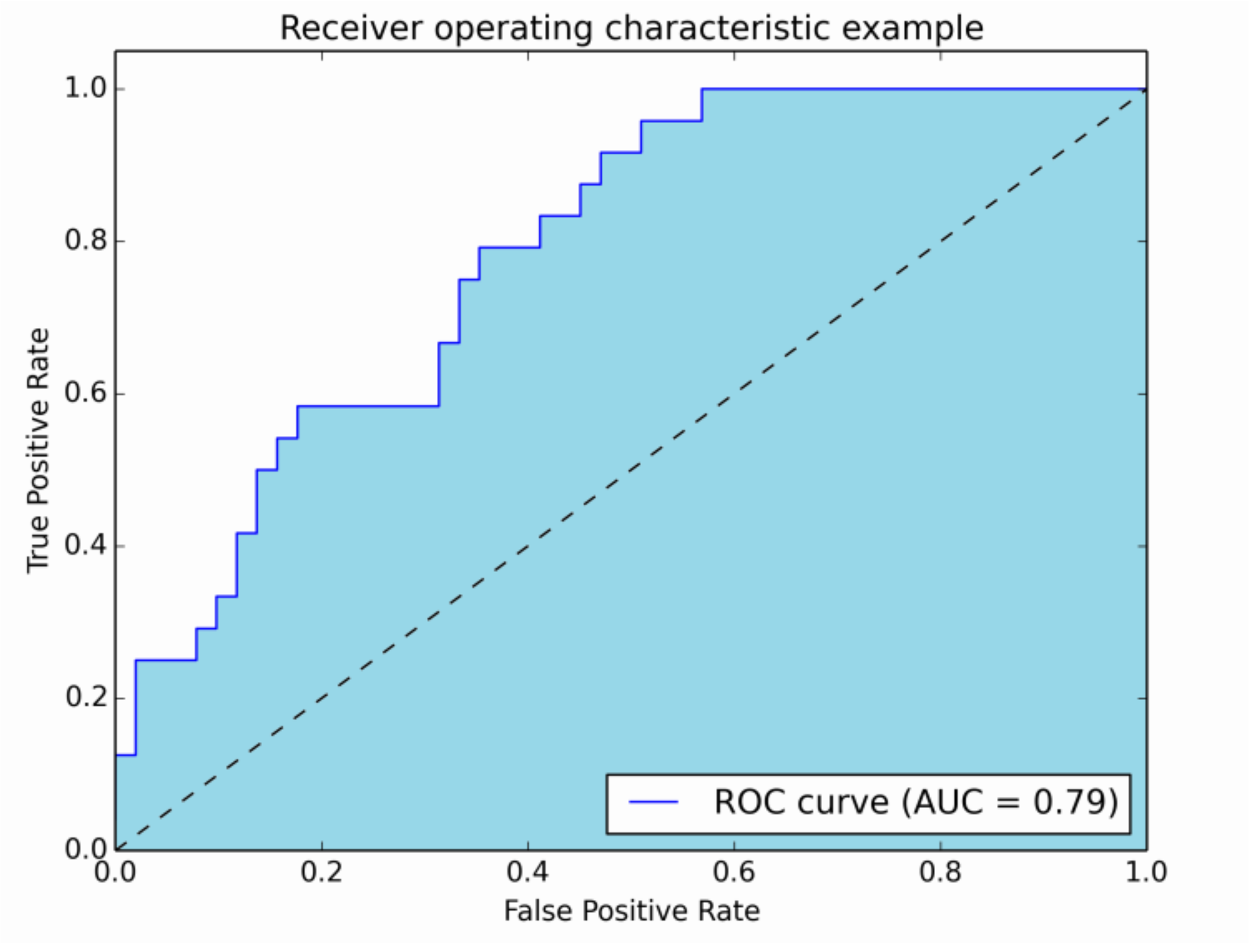
AUC is used most of the time to mean AUROC

- True positive rate (TPR), aka. sensitivity, hit rate, and recall, which is defined as TP/TP+FN. Intuitively this metric corresponds to the proportion of positive data points that are correctly considered as positive, with respect to all positive data points. In other words, the higher TPR, the fewer positive data points we will miss.

- False positive rate (FPR), aka. fall-out, which is defined as FP/FP+TN. Intuitively this metric corresponds to the proportion of negative data points that are mistakenly considered as positive, with respect to all negative data points. In other words, the higher FPR, the more negative data points will be missclassified.

To combine the FPR and the TPR into one single metric, we first compute the two former metrics with many different thresholds (for example 0.00; 0.01, 0.02,… 1.00), then plot them on a single graph, with the FPR values on the abscissa and the TPR values on the ordinate. The resulting curve is called ROC curve, and the metric we consider is the AUC of this curve, which we call AUROC.

The following figure shows the AUROC graphically:



In this figure, the blue area corresponds to the Area Under the curve of the Receiver Operating Characteristic (AUROC). The dashed line in the diagonal we present the ROC curve of a random predictor: it has an AUROC of 0.5. The random predictor is commonly used as a baseline to see whether the model is useful.

Q11 (5 Points) Bootstrap aggregating, also called bagging is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It also reduces variance and helps to avoid overfitting.

Describe the algorithm in detail. What are its downsides?

Solution:

Bagging algorithm

1. Generate m new bootstrap samples

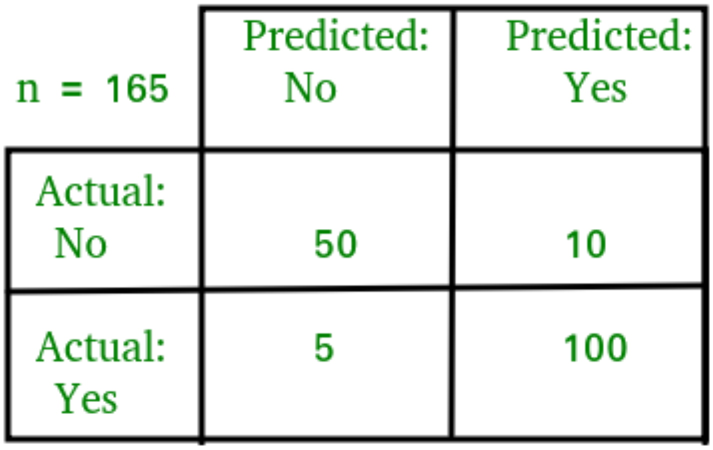
Given a standard training set D of size n, bagging generates m new training sets D_i, each of size n′, by sampling from D uniformly and with replacement.

2. Fit model on each bootstrap sample.
3. Aggregate the m predictions in an ensemble prediction.

Its downside is that one generates many, say m, bootstrap samples and fits them all. If the ensemble prediction isn't much better than a fit on the original training set then the extra computational cost is for nothing.

See https://en.wikipedia.org/wiki/Bootstrap_aggregating

Q12 (5 Points) Given the Confusion Matrix below calculate the accuracy, recall and precision



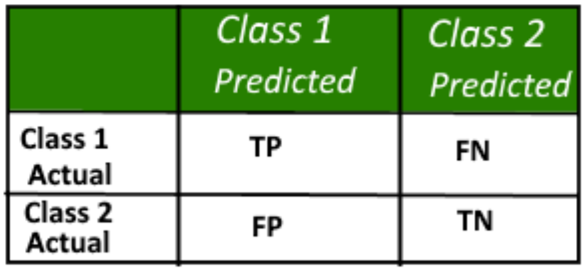
n = 165	Predicted: No	Predicted: Yes
Actual: No	50	10
Actual: Yes	5	100

Solution:

Confusion Matrix see https://en.wikipedia.org/wiki/Confusion_matrix

Just use the equations below.

Confusion Matrix



	Class 1 Predicted	Class 2 Predicted
Class 1 Actual	TP	FN
Class 2 Actual	FP	TN

Accuracy

$$\textbf{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Recall

$$\textbf{\textit{Recall}} = \frac{TP}{TP + FN}$$

Precision

$$\textbf{\textit{Precision}} = \frac{TP}{TP + FP}$$

Q13 (5 Points) What is a loss function? Name three cost functions used in neural networks

Solution:

See https://en.wikipedia.org/wiki/Loss_function

Quadratic cost

Also known as *mean squared error*, *maximum likelihood*, and *sum squared error*, this is defined as:

See https://en.wikipedia.org/wiki/Loss_function#Quadratic_loss_function

Cross-entropy cost

Also known as *Bernoulli negative log-likelihood* and *Binary Cross-Entropy*

See https://en.wikipedia.org/wiki/Cross_entropy

Hellinger distance

https://en.wikipedia.org/wiki/Hellinger_distance

You can find more about this here. This needs to have positive values, and ideally values between 00and 11. The same is true for the following divergences.

Kullback–Leibler divergence

Also known as *Information Divergence*, *Information Gain*, *Relative entropy*, *KLIC*, or *KL Divergence*(See here).

https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence

Itakura–Saito distance

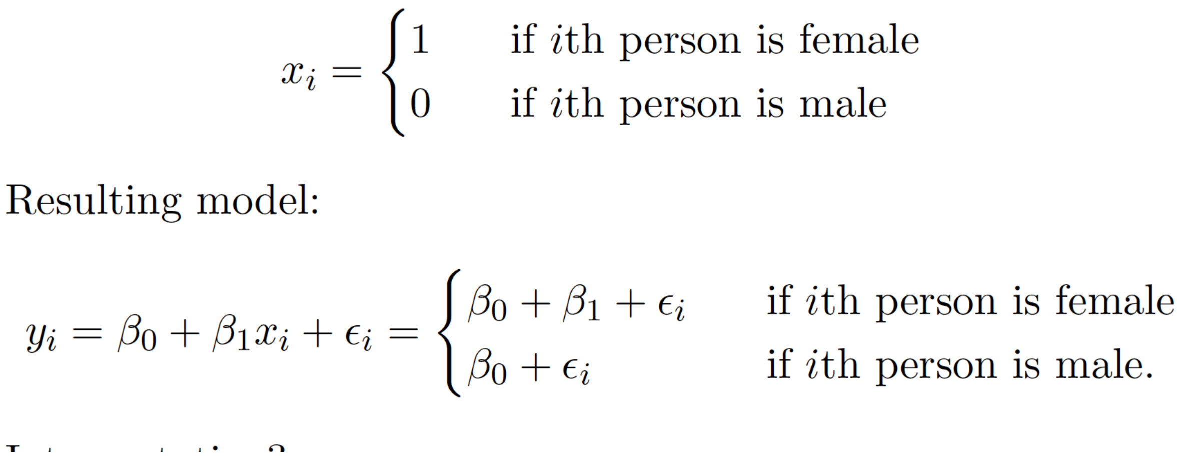
https://en.wikipedia.org/wiki/Itakura%E2%80%93Saito_distance

More loss functions at https://github.com/torch/nn/blob/master/doc/criterion.md

Q14 (5 Points) Assume one wants to use gender (male/female) as an independent variable in regression. How can we encode it?

Solution:

1 female and 0 male (or vice versa)



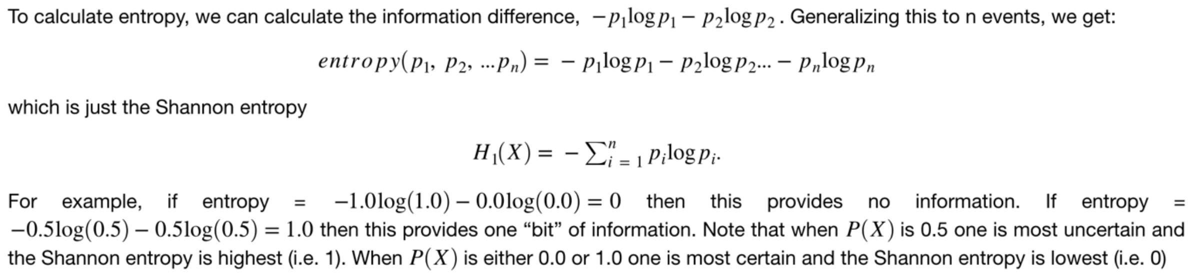
$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male} \end{cases}$$

Resulting model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

Q15 (5 Points) What is the equation for Shannon entropy? What is it a measure of?

Solution:



To calculate entropy, we can calculate the information difference, $-p_1 \log p_1 - p_2 \log p_2$. Generalizing this to n events, we get:

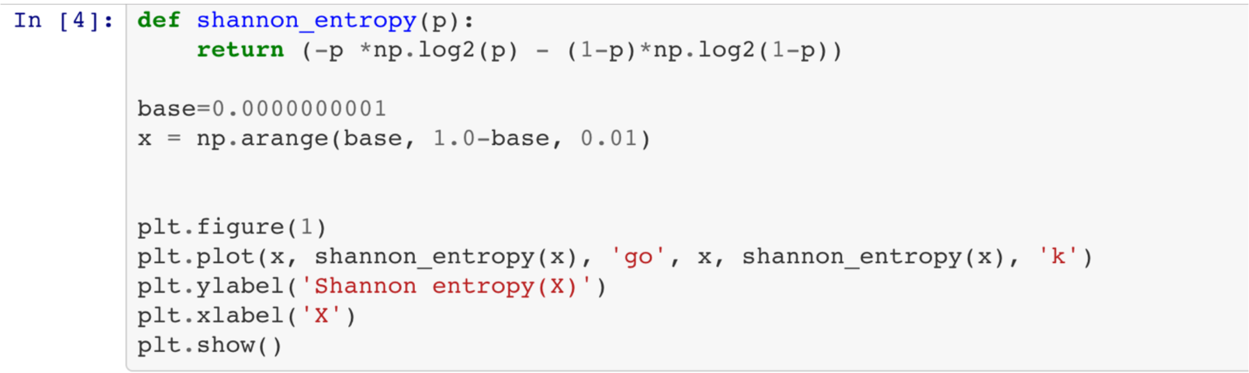
$$entropy(p_1,\ p_2,\ ...p_n) = -\ p_1 \log p_1 - p_2 \log p_2 ... - p_n \log p_n$$

which is just the Shannon entropy

$$H_1(X) = -\sum_{i=1}^{n} p_i \log p_i.$$

For example, if entropy = $-1.0\log(1.0) - 0.0\log(0.0) = 0$ then this provides no information. If entropy = $-0.5\log(0.5) - 0.5\log(0.5) = 1.0$ then this provides one "bit" of information. Note that when $P(X)$ is 0.5 one is most uncertain and the Shannon entropy is highest (i.e. 1). When $P(X)$ is either 0.0 or 1.0 one is most certain and the Shannon entropy is lowest (i.e. 0)

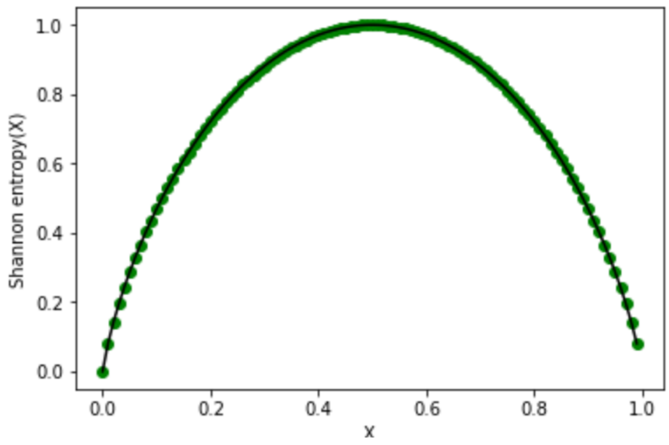
It is a measure of uncertainty/information.



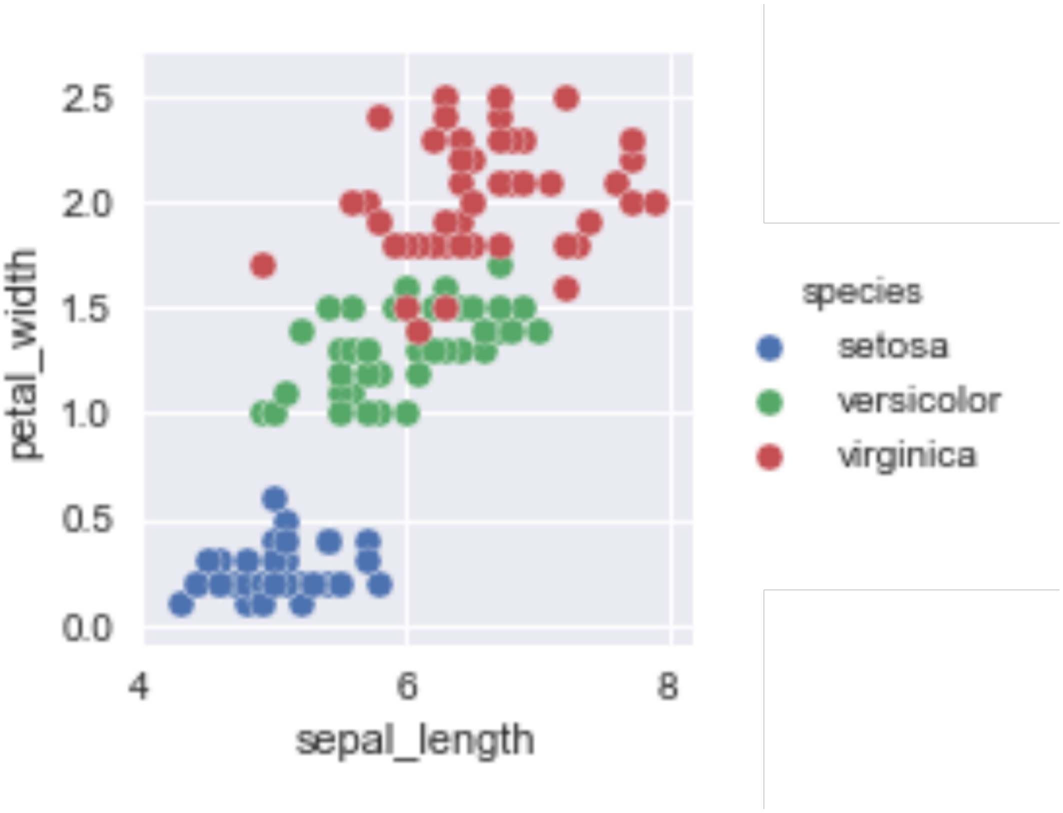
```
In [4]: def shannon_entropy(p):
            return (-p *np.log2(p) - (1-p)*np.log2(1-p))

        base=0.0000000001
        x = np.arange(base, 1.0-base, 0.01)


        plt.figure(1)
        plt.plot(x, shannon_entropy(x), 'go', x, shannon_entropy(x), 'k')
        plt.ylabel('Shannon entropy(X)')
        plt.xlabel('X')
        plt.show()
```



Q16 (5 Points) Below is a scatter plot of petal width versus sepal length in the famous iris data set. If you had to choose either petal width or sepal length to build a classifier for the three species of iris flowers which feature would you use? How well would it perform on this sample data?



Solution:

Clearly petal width is a better predictor. Setosa is perfectly linearly separable from the other two species with a linear discriminant at a petal width of around 0.8 and versicolor and virginica are well separated with a linear discriminant at a petal width of around 1.6.

Q17 (5 Points)

Suppose we have a data set with five predictors, X1 =GPA, X2 = IQ, X3 = Gender (1 for Female and 0 for Male). The response is starting salary after graduation (in thousands of dollars).

Suppose we use least squares to fit the model, and get ˆβ0 = 50, ˆβ1 = 20, ˆβ2 = 0.07, ˆβ3 = 35

(a) Is the following answer correct, and why? For a fixed value of IQ and GPA, males earn more on average than females.

(b) Predict the salary of a female with IQ of 110 and a GPA of 4.0.

(c) True or false: Since the coefficient for the IQ term is very small, there is very little evidence of an IQ effect. Justify your answer.

Solution:

a) Is the following answer correct, and why?

For a fixed value of IQ and GPA, males earn more on average
than females.

Males earn less.

The same values of IQ and GPA generate the same values for males and females however the gender variable ˆβ3 = 35 is positive and 1 means female so if female the prediction will be 35 more for a fixed value of IQ and GPA.

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male} \end{cases}$$

Resulting model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

(b) Predict the salary of a female with IQ of 110 and a GPA of 4.0.

X1 =GPA, X2 = IQ, X3 = Gender (1 for Female and 0 for Male)

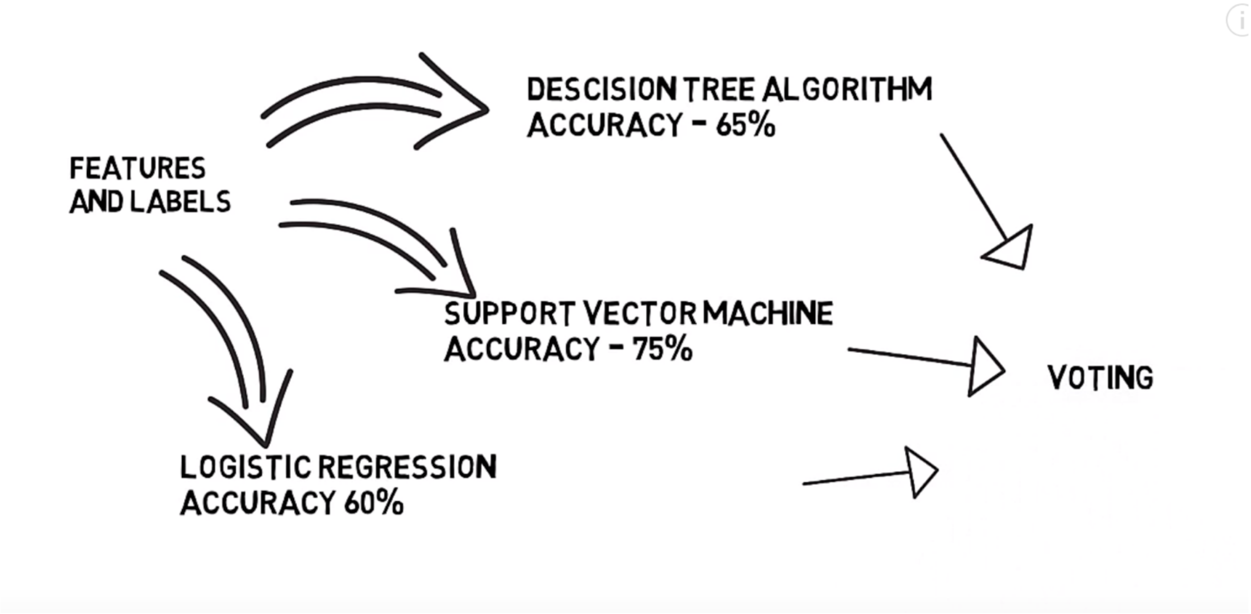
ˆβ0 = 50, ˆβ1 = 20, ˆβ2 = 0.07, ˆβ3 = 35

Use equation y = β0 + β1* GPA + β2* IQ + β3* Gender

(c) True or false: Since the coefficient for the IQ term is very small, there is very little evidence of an IQ effect. Justify your answer.

False. We care about the p-value not the coefficient.

Q18 (5 Points)

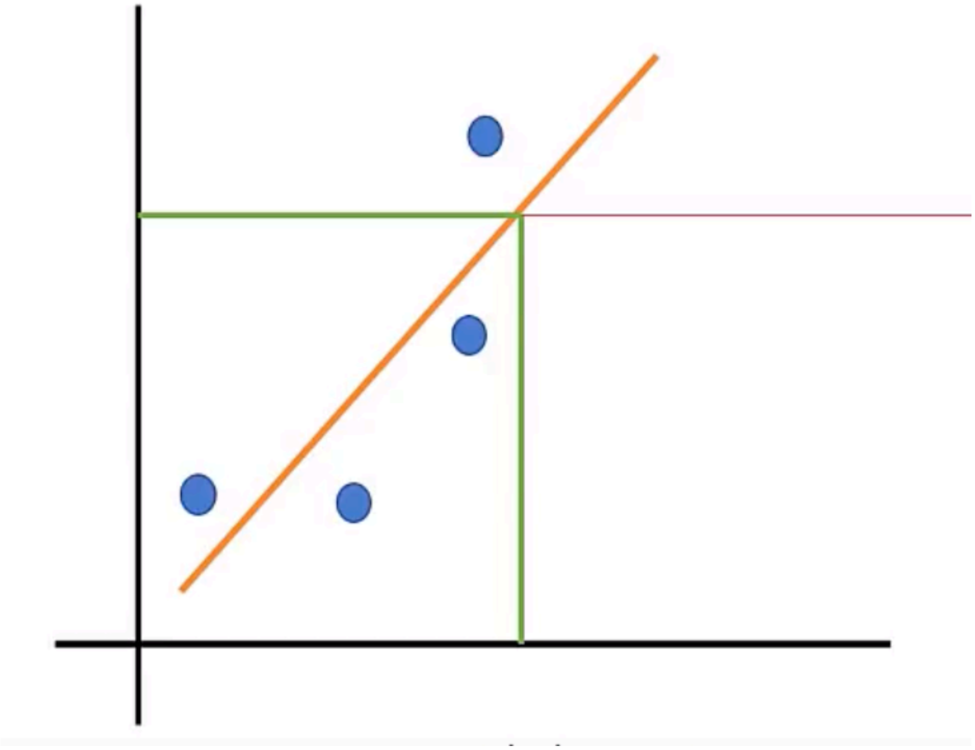


You combine three different algorithms, decision tree, SVM and logistic regression in to one predictor. What is this technique called? How could one combine the three different algorithms in to one prediction?

Solution:

This is called "stacking." There are many methods of aggregation; but taking the mean for regression and majority vote for classification are the simplest.

Q19 (5 Points)



Is linear regression a high or low bias model? Why?

Solution:

Linear regression is a high bias model due to its limited flexibility to fit the residuals in the training data.

Q20 (5 Points)

Describe the cross-validation algorithm. Use pseudo-code and include enough detail to implement it in code.

What are cross-validation's advantages and disadvantages?

Solution:

k-folds

Split randomly selects a (fixed) k number of roughly equal sets without replacement.

This could be done by flipping a k-sided coin and using that to distribute the data in to k bins.

k-training rounds

Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k − 1 subsamples are used as training data. The cross-validation process is then repeated k times, with each of the k subsamples used exactly once as the validation data.

Combining rounds

The k results can then be averaged to produce a single estimation.

Advantages?

The advantage of this method over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation exactly once. 5 and 10-fold cross-validation is commonly used, but in general k remains an unfixed parameter.

Another advantage of this method is that it matters less how the data gets divided. Every data point gets to be in a test set exactly once, and gets to be in a training set k-1 times. The variance of the resulting estimate is reduced as k is increased.

Disadvantages?

The disadvantage of this method is that the training algorithm has to be rerun from scratch k times, which means it takes k times as much computation to make an evaluation. A variant of this method is to randomly divide the data into a test and training set k different times. The advantage of doing this is that you can independently choose how large each test set is and how many trials you average over.