**Supervised Learning**

**Learning Objectives**

* Develop an understanding of supervised learning and its common applications
* Be able to perform regression and classification techniques to solve real-world problems

**Work to Complete**

In this unit, you'll:

* Complete a logistic regression case study
* Complete a Decision Trees case study
* Complete a Random Forest case study
* Complete a Gradient Boosting case study

Supervised Learning is the bread and butter of machine learning. Consider a teacher facilitating a student’s learning process by showing them what to do and then having them do that task repeatedly. Supervised learning is similar — a data scientist uses a training dataset to teach an algorithm how to use a mapping function from an input to an output. The algorithm learns to approximate the mapping function so well that the algorithm can predict output variables for that data. Put another way — you’ll use supervised learning when you give your machine labeled training data and encode procedures for the machine to learn to assign those labels itself.

Let’s take a look at a simple example. If you wanted to predict if a cat is a particular breed, you could load up a ton of cat information, such as height, body hair length, nose color, and more. These characteristics are called “features.” All of these features combined make up the training data that your algorithm can draw from — the algorithm can predict a cat’s breed based on the cat’s height, body hair length, nose color, etc.

This unit will explore several supervised learning methods, including logistic regression, decision trees, random forest, gradient descent, series analysis, and forecasting.

### Logistic Regression

Logistic regression is a technique borrowed from the world of statistics — data scientists use it when working on classifications with two class values (aka binary classification problems). A logistic regression model analyzes the relationship between independent variables to predict a dependent data variable. Work through the resources found in this subunit to learn more.

**Quizes**

What was the moral of the ‘Millennials living at home’ story?

a. Statistics don’t lie

**b. Always check what the variables mean**

c. Always do machine learning properly

d. Don’t gather too much data

Suppose a survey on the quality of life of single mothers is launched, and a much lower number of single mothers than expected actually take the survey due to the tasks and pressures associated with being a single mother. Nonetheless, the local paper misleadingly publishes an article stating that single mothers have an easy and quiet life (because the ones that did that the survey reported this). What’s the name of the active statistical bias here?

a. Length bias

b**. Response bias**

c. Gender bias

d. Confirmation bias

In the example of the 1936 Landon vs FDR presidential election, why did Literary Digest get things so wrong with such a large sample size of over 2 million?

a. They used imprecise or vague columns

b. 2 million was still not enough

c. They set up their confidence interval and significance level incorrectly

**d. Their sample method was poor**

The revelation regarding the armor placement on planes was…

**a. To put the armor on the parts which hadn’t been struck on the planes that successfully came back**

b. To put the armor on the most frequently struck parts

c. To put the armor on the most crucial and delicate parts of the plane, regardless of whether they were struck

d. None of the above

Why did students have such low life expectancy in Lombard’s study of 1835?

a. Students party all the time and drink too much

**b. Studenthood is transient - alas!**

c. The other jobs were ones that make you live very long

d. None of the above

What causes the particular instance of the class size paradox that was discussed in the video?

a. Statistics is illogical

**b. Perspective matters: depending on your vantage point and how you set up your calculations, results differ**

c. One of the calculations is correct and the other incorrect: the class size ratio is very good in the college

d. One of the calculations is correct and the other incorrect: the class size ratio is very poor in the college

The reported statistic that 21% of prison inmates are serving 20 years or more is explained by the fact that a given time-slice is more likely to pick out inmates serving longer sentences

**TRUE**

FALSE

What is the bias of an estimator, statistically speaking?

**a. How far off from the truth that estimator is, on average**

b. How affected that estimator is by prejudices and value-judgments, regardless of whether it’s far from the truth or not

c. How far off from the truth that estimator is, at its worst

d. How far off from the truth that estimator is, due to prejudices and value-judgments

We cannot have a low Mean-Squared Error (MSE) unless we have 0 bias

TRUE

**FALSE**

Extra: Bias-variance decomposition forms the basis of regression regularization methods such as Lasso

**TRUE**

FALSE

In the Unbiased estimation: Poisson example section, it’s asserted that (-1)X is a good estimator of e-2λ

TRUE

**FALSE**

What does Fisher weighting try to achieve?

**a. The combination of independent, unbiased estimators for a parameter into one estimator**

b. The separation of mutually dependent, biased estimators into independent, unbiased ones

c. The separation of independent, biased estimators into independent, unbiased ones

d. None of the above

With the Fisher weighting method, higher variance means lower reliability

**TRUE**

FALSE

What’s the optimal way to relate the weight value to the variance?

a.Make the weight directly proportional to the variance

**b. Make the weight inversely proportional to the variance**

c. Make the weight and the variance equal

d. None of the above

One of the issues with Fisher Weighting is that, in practice, we often have several different datasets we want to combine that are both biased and dependent

**TRUE**

FALSE

Statistician Nate Silver, by contrast with other analysts, concentrated on a number of factors in his model. Which of the following was NOT a factor claimed by the lecturer to have been accounted for by Silver?

a. Where the data came from

b. How the data were collected

c. How the data should be combined

**d. All logically possible ways in which the data could have been combined**

Statistician Nate Silver, by contrast with other analysts, concentrated on a number of factors in his model. Which of the following was NOT a factor claimed by the lecturer to have been accounted for by Silver?

a. Where the data came from

b. How the data were collected

c. How the data should be combined

d. All logically possible ways in which the data could have been combined

Silver’s aim was to come up with a system that represented all of these factors as weights, and then use those in a predictive model

**TRUE**

FALSE

What is the Bonferroni Correction and when is it used?

a. It’s a reduction of the significance level, used when the Null is being incorrectly rejected (False Negative)

**b. The division of the significance level by the number of hypotheses being tested, used when we’re testing multiple hypotheses**

c. The multiplication of the significance level by the number of hypotheses being tested, used when we’re testing multiple hypotheses

d. An increase in the significance level, used when the Null is being incorrectly accepted (False Positive)

Suppose we have a dataset that has lots of variables, and we’ve looked at many combinations of those variables. It’s not likely that some of those variables are correlated

TRUE

**FALSE**

The p-value of an observation is roughly the probability of seeing a result at least as extreme as that observation, on the assumption of the Null hypothesis

**TRUE**

FALSE

The regression line gives the best prediction of y from x, in the sense of minimizing the sum of squared residuals

**TRUE**

FALSE

Which of the following are NOT examples mentioned in the lecture of Regression Toward the Mean (= RTTM):

a.Test scores

**b. Personalities**

c. Inherited characteristics

d. Traffic accident locations

With the instance of the regression paradox mentioned in the lecture, y = the child’s height, x = the parent’s height, and the variables have been standardized. What does it mean to say those variables have been standardized?

a. They have Standard Deviation (= SD) 0 and mean 1

b. They have SD 0 and mean 0

**c. They have SD 1 and mean 0**

d. They have SD 1 and mean 1

Continuing with the regression paradox, if the regression line is y = rx, then, mathematically speaking, what should x equate to? (The symbols ‘/’ and ‘\*’ mean division and multiplication respectively). Note: the question is not asking for what x does equal to, but what it initially seems x should equate to.

**a. x = 1/r \* y**

b. x = 1/y \* r

c. x = r \* y

d. x = r/y

In statistical fact, what does x equate to?

a. None of the below

b. x = 1/r \* y

c. r

**d. x = ry**

What is the best explanation of the regression paradox?

a. The illogicality of the universe

**b. Regression Toward the Mean**

c. Statistics uses different axioms to ordinary mathematics

d. None of the above

If we’re using x to predict y linearly, what are we hoping?

**a. That y is approximately a linear function of x**

b. That x is not in fact squared (making the equation quadratic)

c. That y is not in fact squared (making the equation quadratic)

d. None of the above

Regression can be represented as follows: y = Xβ + ε In general, we’re trying to find the function of X that does the best job at predicting y. This function could be linear or nonlinear. But if we have a normal distribution, then the best predictor of y as a function of X is:

a. Quadratic

Nonlinear

**Linear**

Neither linear nor nonlinear

What’s the problem with R-squared as a metric?

**a. It doesn’t quantify the predictive power**

b. It’s hard to calculate

c. It doesn’t measure the goodness of fit

d. It’s imprecise

‘Heteroscedasticity’ refers to a model such that the variability of a variable is equal across the range of values of a second variable predicting it

TRUE

**FALSE**

With the NYC Housing Example, why was the distribution bimodal?

a. Because the data were collected inappropriately

b. Because New York is more expensive than other cities

**c. Because Manhattan is more expensive than other parts of New York**

d. Because Brooklyn is cheaper than other parts of New York

You can take the log of a negative number

TRUE

**FALSE**

With logistic regression:

a. The response variable y is continuous

**b. The response variable y is categorical**

c. The response variable y is neither continuous nor categorical

d. The response variable y is equal to the predictor variable x

Is a big table a bad way of summarizing the results of linear regression analysis? If so why?

**a. Yes, it’s overwhelming and over-emphasizes the p-value**

b. Yes, it’s underwhelming and under-emphasizes the p-value

c. No, plotting the results is less clear

d. No, a big table contains all the information we need

Collinearity occurs when

a. The response variable is correlated with itself

**b. The predictor variables are correlated with themselves**

c. The predictor variables are correlated with the response variable

d. None of the above

Standardizing the variables means giving them mean 0 and standard deviation 1. What is a benefit, and drawback, of doing this, respectively?

a. Increases interpretability, but makes them no longer on an equal footing

b. Makes the variables plottable, but can decrease interpretability

**c. Makes the variables on an equal footing, but can decrease interpretability**

d. Increases interpretability, but distorts the truth

If someone’s probability of experiencing an outcome is p, then the person’s odds of the outcome are:

a. p2

**b. p/(1-p)**

c. (1-p)/p

d. None of the above

Suppose we’re investigating a link between variables X and Y. What is a confounding factor Z?

**a. Z is a variable correlated with both X and Y that may explain that link**

b. Z is a variable that’s correlated with neither X nor Y but is confusing in itself

c. Z is a variable that only Y is correlated with

d. Z is a variable that only X is correlated with

What does Logit(p) equate to?

**a. The log(p/1-p)**

b. The Maximum Likelihood Estimation

c. The log(1-p/p)

d. None of the above

What is the aim of Maximum Likelihood Estimation (MLE)?

a. To get parameter values that make the actual data is unlikely as possible

b. To get parameter values that make the model as easy to apply as possible

**c. To get parameter values that make the actual data is likely as possible**

d. None of the above

With the geometric example given by the lecturer, what was the point supposed to illustrate?

**a. Our intuitions become less accurate as dimensionality increases**

b. Our intuitions become more accurate as dimensionality increases

c. Dimensionality is just a blessing

d. Dimensionality is just a curse

Do nearest neighbor methods become harder as the dimensions increase? If so, why?

a. No, it makes no difference

b. No, the increase in complexity is negligible

**c. Yes, we have trouble finding close-by neighbors**

d. Yes, we have trouble choosing between so many neighbors

What is the tall vs wide data distinction?

a. Tall data has many variables, wide has many individuals

**b. Tall data has many individuals, wide data has many variables**

c. Tall data has only a few variables

d. Wide data has only a few individuals

What, in essence, is Ridge Regression?

a. Linear regression, but where the response variable y is categorical

**b. An extension of linear regression such that we prevent the sum of squared β’s from getting too high with a penalty term**

c. An extension of linear regression such that we prevent the sum of absolute values from getting too high with a penalty term

d. None of the above

What is the strange and fascinating general lesson of Stein’s paradox, as explained by the lecturer?

a. We can estimate a vector of means only if we’ve pre-determined a loss function, and such a function is always admissible.

b. The Mean-Squared error is an inadmissible estimator.

**c. In certain cases, it is inadmissible to solve apparently separate problems separately; in those cases, we’d do better across the board if we combined apparently independent problems.**

d. None of the above

The absolute value function is:

a. An S-shape

**b. A V-shape**

c. A diamond shape

d. A circle shape

What is the distinctive motivation for LASSO over Ridge?

a. LASSO doesn’t induce sparsity

**b. LASSO induces sparsity**

c.LASSO is computationally feasible, unlike Ridge

d. LASSO can be cast into a convex optimization framework

Which of the following is a function performed by Ridge Regression for the data scientist?

**a. Shrinkage**

b. Clustering

c. Time-series Analysis

d. Logistic regression

LASSO is distinctively useful in cases where:

a. Traditional linear regression is also useful

b. We have more individuals than variables

c. Ridge is useful

**d. We have more variables than individuals**

Why do we get a diamond, not a circle, when plotting LASSO?

a. Because with LASSO we’re looking at points where the sum of absolute values is greater than a constant

**b. Because with LASSO we’re looking at points where the sum of absolute values is less than a constant**

c. Because with LASSO we’re trying to induce sparsity

d. Because with LASSO we’re not trying to induce sparsity

### Decision Trees

Tree-based algorithms, like decision trees and random forests, are some of the most popular and effective classification and regression algorithms, especially when dealing with complex datasets.

A decision tree is a flowchart-like tree structure; each internal node represents a test on an attribute, each branch represents one of the outcomes of that test, and each leaf node represents a class label. You’ll get a chance to create a decision tree when you work on a case study at the end of this subunit.

Please watch from 3:30 - 37:34. Tree-based algorithms (e.g. decision trees and random forests) are some of the most popular and effective classification and regression algorithms, especially when working with complex datasets. In this Harvard University lecture, you’ll learn about decision trees (5:00) and understand ensemble methods, specifically starting with bagging (1:00). By the end of the lecture, you’ll be able to summarize how decision trees work, which is an important question that comes up during job interviews. View the presentation slides [here](https://github.com/cs109/2015/blob/master/Lectures/11-DecisionTreesAndRandomForest.pdf).

Which of the following is NOT a major benefit of a decision tree model?

a. Fast training

b. Easy to interpret

c. Fast prediction

**d. Predictive power**

What, ultimately, are we after when constructing a decision tree model?

a. Interpretation of the data

b. A more precise rendering of the data

**c. A decision boundary**

d. Explanation of the past

With a binary decision tree, asking a single question…

a. Splits the feature space into four

b. Splits the feature space into eighths

**c. Splits the feature space in half**

d. Splits the feature space into sixteenths

A decision tree partitions the feature space into cells. What are these cells?

a. Our data points

b. Spaces containing one data point each

c. Our features

**d. Our categories or classes**

The partitioning of the feature space by a binary decision tree is always a diagonal line

TRUE

**FALSE**

Which of the following is NOT a criterion for identifying the purity of a node?

a. Entropy

b. Gini index

c. Misclassification index

**d. Depth**

What’s the aim of looking at node purity gain?

**a. To work out how pure splitting on that feature and threshold is**

b. To work out how deep the tree should be

c. To work out how pure the tree is

d. None of the above

Consider the following algorithm. i. Check if already finished ii. For each feature xi: a. Calculate the gain from splitting on xi b. Let xbest be the feature with the lowest gain iii. Create a decision node that splits on xbest Which step is mistaken?

a. (i)

b. (ii)(a)

**c. (ii)(b)**

d. (iii)

The value for misclassification and gini gain can come apart, for a given node

**TRUE**

FALSE

Which of the following is not a relevant consideration mentioned by the lecturer for when the decision tree algorithm should terminate?

**a. The tree is too shallow**

b. node contains less than x data points

c. max depth is reached

d. max purity is sufficient

How does extreme overfitting typically manifest itself in a decision tree?

a. A very deep tree

b. A tree with high impurity

**c. A little cell around every data point in the set**

d. A very shallow tree

What is the way of treating overfitting?

**a. Pruning**

b. Using entropy

c. Using an SVM instead

d. Comparing pre- and post-split entropy

Which of the following is NOT a disadvantage of decision trees:

a. Overfitting

**b. Slow execution**

c. Sensitive to small changes in the data

d. Only axis-aligned splits

Let’s compare decision trees and SVMs. SVMs are better at handling missing values

TRUE

**FALSE**

Decision trees are better at extracting linear combinations of features

TRUE

**FALSE**

What is the main purpose of ensemble methods?

**a. To augment the predictive power of decision trees**

b. To make decision trees execute faster

c. To make decision trees better at handling missing values

d. To make decision trees better at extracting linear combinations of features

With ensemble methods such as random forest, what’s the name of the method we use to ensure that all the trees in the ensemble don’t just uselessly learn in exactly the same way?

a. Permutation test

b. SVM

c. Cross-validation

**d. Bootstrap**

Which of the following claims is NOT true about bootstrapping?

a. It’s a resampling method for statistics

**b. It’s not a useful way to get the error bars on estimates**

c. It draws with replacement

d. It’s used to make the trees in an ensemble method learn in different ways

Bagging

**a. Normally uses one type of classifier**

b. Is an ensemble method where you fit many models to all your data.

c. Generates models that are easy to interpret.

We should use simple bootstrap to generate train and test data from the same dataset

TRUE

**FALSE**

### Ensemble Methods and Random Forest

Put a bunch of decision trees together, and what do you get? The random forest. The random forest is a classification algorithm that’s made up of decision trees. You’ll use random forest when working on classification and regression tasks.

How do you select which of the learning algorithms to use? It’s often said that people like having their cake and eating it, too. The machine learning equivalent of that proverb is ensemble methods, which is a learning algorithm that constructs a set of classifiers and to identify new data points using a combination of predictions. Ensembles tend to be more accurate and robust than single classifiers. This Harvard University lecture starts with random forests (8:00), which is essentially bagging for decision trees. The talk then covers a different approach to ensemble learning called boosting (25:00), specifically a technique called AdaBoost (30:00). View the presentation slides [here](https://github.com/cs109/2015/blob/master/Lectures/12-Ensemble%20Learning%20and%20Random%20Forests.pdf).

**Ensemble Method Quizzes**

Ensemble methods:

a. Combine many learners, provided each is individually strong

**b. Combine many learners, each of which could be individually weak**

c. Come in just one variety

d. None of the above

Which of the following is not an ensemble method?

a. Random forest

b. Bagging

c. Boosting

**d. Clustering**

In the movie recommendations analogy to random forest, what was our leaving out certain films we like/dislike to one friend, but leaving out other films to a different friend, the analog of?

a. Bagging

b. Gradient descent

**c. Bootstrapping**

d. Entropy

Suppose in the movie recommendation analogy I like both Wild at Heart and Blue Velvet, but not because they’re directed by David Lynch. How is this accounted for in random forest?

a. By pruning the tree

**b. By making randomness partly determine which attribute to split on at each node**

c. By overfitting

d. By gradient descent

Why is the wisdom of crowds dubious?

**a. Because there’s no reason to suspect that positive bias always counterbalances negative bias**

b. Because there’s no reason to suspect that there is a General Will

c. Because human beings’ beliefs converge to falsehood, due to falsifying cognitive faculties

d. None of the above

The principal difference between boosting and bagging is that with boosting, weak learners evolve over time, and votes are weighted

**TRUE**

FALSE

Which of the following is NOT a tuning parameter for boosting:

a. Number of trees

b. Number of splits in each tree (often stumps work well)

c. Parameters controlling how weights evolve

**d. Splitting criterion**

Consider the AdaBoost reweighting step.

**a. If the indicator is 0, then we’re not changing the weight. If the indicator is 1, then we misclassified the observation, so we put more weight on it.**

b. If the indicator is 0, then we misclassified the observation, so we put more weight on it. If the indicator is 1, then we’re not changing the weight.

c. If the indicator is 0, then we misclassified the observation, so we square its weight. If the indicator is 1, then we’re not changing the weight.

d. None of the above.

Boosting can be prone to overfitting .

**TRUE**

FALSE

A complicated decision tree can give you complex boundaries but will overfit. The bagging process applied properly will smooth it out.

**TRUE**

FALSE

**Case Study - Random Forest**

2 - 4 Hours

55 Points

Data science plays a vital role in how we understand and react to real-world situations. It can help us understand the likelihood of an event occurring and inform decisions about how to respond to that event. Being able to understand data is particularly important in the fields of health and science. In this case study, you'll use Random Forest and logistic regression to understand the scope of the Coronavirus using data from December and January of 2020. This case study is an excellent example of how data scientists can help share crucial insights about occurrences that have an impact around the world.