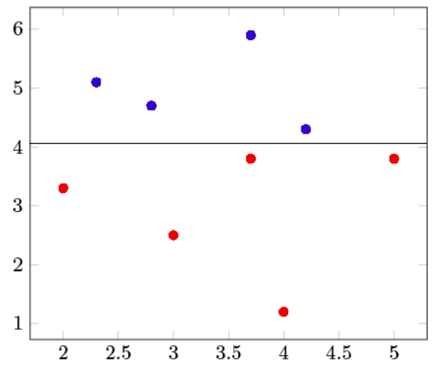
Knowledge Checks for edX 6501x

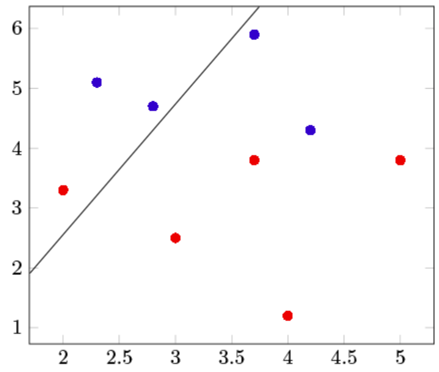
2. Classification

Knowledge Check 2.1: Introduction to Classification

Graph A



Graph B



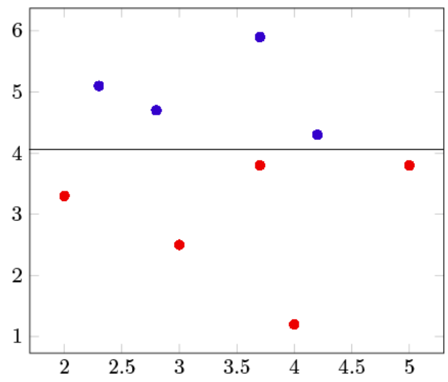
1. Which graph shows a classifier that separates between the blue and red points?

- A
    - All blue points are above the line and all red points are below the line
- B
    - Blue and red points are on the same side of the line

2. In Graph A, which color would your classifier predict for a point in the lower right-hand corner of the graph?

- Blue
    - Below the line there are only red points
- Red
    - Below the line there are only red points

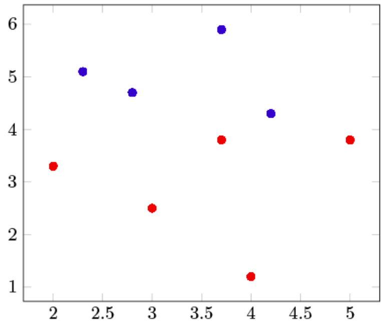
Knowledge Check 2.2: Choosing a Classifier



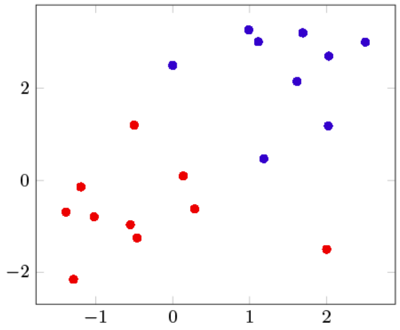
Which factor is most important for classification?

- Factor on the horizontal axis
    - ○ The classifier is about the same for any horizontal-axis value
- Factor on the vertical axis
    - ○ Almost all that matters is whether the vertical-axis value is above or below a specific value
- Both
    - ○ The classifier is about the same for any horizontal-axis value

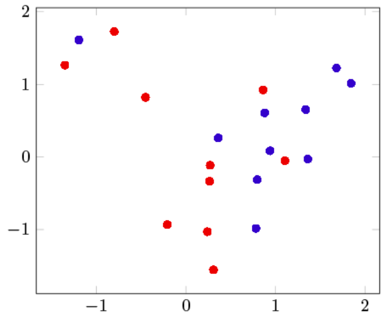
Graph A



Graph B



Graph C



Which graph requires a soft classifier, because no hard linear classifier exists for it?

- A
    - ○ A horizontal line between the blue and red points can be a hard classifier
- B
    - ○ A diagonal line from upper left to lower right can separate the blue and red points
- C
    - ○ It is impossible to draw a line separating the blue and red points

Knowledge Check 2.3: Data Definitions

A survey of 25 people recorded each person’s family size and type of car.

1. Which of these is a data point?
    - The 14th person’s family size and car type
        - A data point is all the information about one observation
    - The 14th person’s family size
        - A data point is all the information about one observation, not a single attribute value of one observation
    - The car type of each person
        - A data point is all the information about one observation, not all the values of one attribute.

2. Which of these is structured data?
    - The contents of a person’s Twitter feed
        - Free-text responses are not structured
    - The amount of money in a person’s bank account
        - Every entry will be a number of dollars and cents
3. Which of these is time series data?
    - The average cost of a house in the United States every year since 1820
        - The same thing is measured at yearly time intervals
    - The height of each professional basketball player in the NBA at the start of the season
        - Different things are being measured, all at the same instant in time

Knowledge Check 2.4: Support Vector Machines (SVM)

Which of these two terms measures the error in classifying all of the data points?

- $\sum_{i=1}^{m}(a_i)^2$ ✖ Minimizing this term helps maximize the margin between the two categories of points.

- $\sum_{j=1}^{n} max\{0, 1-(\sum_{i=1}^{m} a_i x_{ij} + a_0)y_j\}$ ✓ This term measures classification error

Knowledge Check 2.6: Advanced SVM

Look at the classification error expression below. For which set of data points (1-20 or 21-50) is it more important to avoid classification errors?

$$\sum_{j=1}^{20} 5 \times max\{0, 1-(\sum_{i=1}^{m} a_i x_{ij} + a_0)y_j\}$$
$$+\sum_{j=21}^{50} 200 \times max\{0, 1-(\sum_{i=1}^{m} a_i x_{ij} + a_0)y_j\}$$

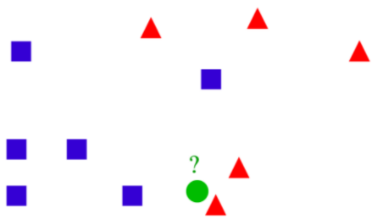
- 1-20
    - The multiplier for classification errors is only 5 for data points 1-20, and is 200 for data points 21-50
- 21-50
    - The multiplier for classification errors is 200 for data points 21-50, much more than 5 for data points 1-20

Knowledge Check 2.7: Scaling and Standardization

Which set of data is scaled between 0 and 1?

- {5, 12, 27, 29}
    - The values are not between 0 and 1.
- {0.0, 0.2, 0.6, 1.0}
    - The values range from 0 to 1
- {0.3, 0.4, 0.7, 0.75}
    - The values are between 0 and 1, but do not span the full range.

Knowledge Check 2.8: K-Nearest Neighbor Classification



What color the k-nearest-neighbor classification algorithm suggest for the green point when k=3?

- Blue
- Red
    - The three closest points to the green one are two reds and one blue. There are more reds than blues

3. Validation

Knowledge Check 3.1: Introduction to Validation

If we use the same data to fit a model as we do to estimate how good it is, what is likely to happen?

- The model will appear to be better than it really is.
  - The model will be fit to both real and random patterns in the data. The model's effectiveness on this data set will include both types of patterns, but its true effectiveness on other data sets (with different random patterns) will only include the real patterns
- The model will appear to be worse than it really is.
- The model will appear to be just as good as it really is.

Knowledge Check 3.2: Validation and Test Data Sets

When comparing models, if we use the same data to pick the best model as we do to estimate how good the best one is, what is likely to happen?

- The model will appear to be better than it really is.
  - The model with the highest measured performance is likely to be both good and lucky in its fit to random patterns.
- The model will appear to be worse than it really is.
- The model will appear to be just as good as it really is.

Knowledge Check 3.3: Splitting Data

Which should we use most of the data for: training, validation, or test?

- Training
  - Most experts recommend using 50-70% of the data for training, and splitting the rest equally between validation and test.
- Validation
- Test

Knowledge Check 3.4: Cross-Validation

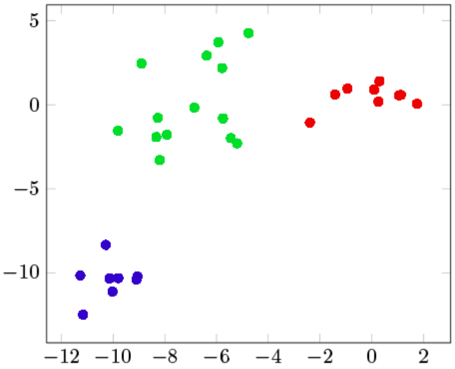
In k-fold cross-validation, how many times is each part of the data used for training, and for validation?

- k times for training, and k times for validation
  - Each of the k times the model is fit, every part of the data is used exactly once.
- 1 time for training, and k-1 times for validation
  - Each of the k times the model is fit, most of the data is used for training.
- k-1 times for training, and 1 time for validation
  - Each of the k times the model is fit, a different part of the data is used for validation and the rest is used for training.

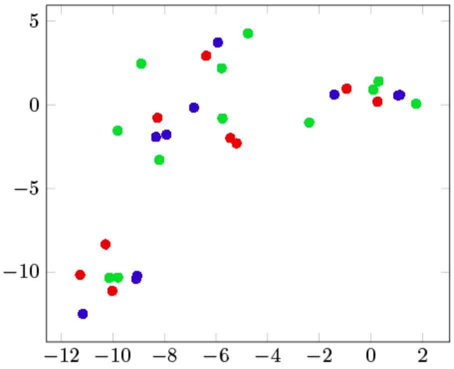
4. Clustering

Knowledge Check 4.1: Introduction to Clustering

Graph A



Graph B



For which of these two graphs does the coloring show a good clustering into 3 clusters?

- A
  - Each color’s nodes are grouped close to each other.
- B
  - In a clustering, nodes should be clustered into groups of nodes that are all close to each other.

Knowledge Check 4.2: Distance Norms

Straight-line distance corresponds to which distance metric?

$$\sqrt[2]{\sum_{i=1}^{n}(x_i - y_i)^2}$$

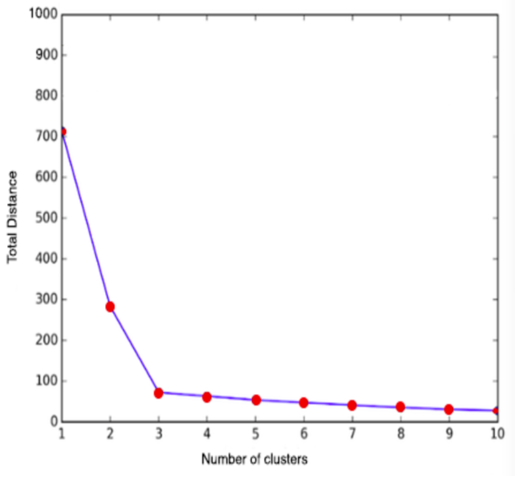
- 1-norm
- 2-norm
  - The power and root are the same as the norm.
- ∞ norm

Knowledge Check 4.3: K-means Clustering

The k-means algorithm for clustering is a “heuristic” because…

- …it runs quickly.
    - Some algorithms run quickly and are guaranteed to find the optimal solution, and some heuristics can be slow.
- …it never gets the best answer.
    - Heuristic algorithms might sometimes get the best answer, but they're not guaranteed to.
- …it isn't guaranteed to get the best answer.
    - Heuristic algorithms are not guaranteed to find the best answer.

Knowledge Check 4.4: Practical Details for K-Means

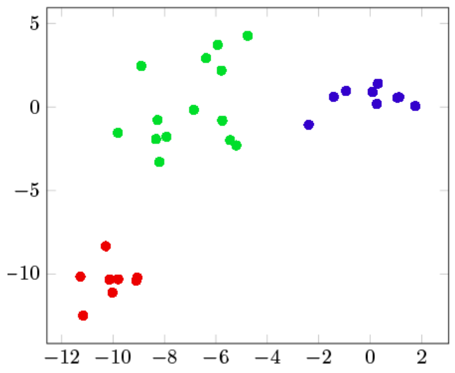


x

Based on the elbow diagram above, how many clusters would you recommend using?

- 1
    - Total distance from points to cluster centers is much higher with 1 cluster.
- 2
    - Significant improvement still exists beyond 2 clusters.
- 3
    - Cluster 3 makes a significant improvement, but additional clusters do not.
- 4
    - 4 clusters do not improve significantly over 3 clusters

Knowledge Check 4.5: Clustering for Prediction



In the figure above, what color would the clustering model predict for a point in the upper-right corner?

- Red
  - The new point would be far from the red points.
- Blue
  - The blue points are closest to the new point.
- Green
  - The green points are far from the new point.

Knowledge Check 4.6: Supervised vs Unsupervised Learning

1. A group of astronomers has a set of long-exposure CCD images of various distant objects. They do not know yet which types of object each one is, and would like your help using analytics to determine which ones look similar. Which is more appropriate: classification or clustering?

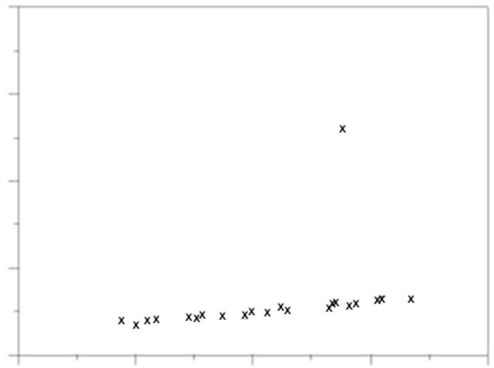
- Classification
- Clustering
  - Without knowing what each image is, the best you can do is group the images into ones that are similar (unsupervised learning).

2. Suppose one astronomer has categorized hundreds of the images by hand, and now wants your help using analytics to automatically determine which category each new image belongs to. Which is more appropriate: classification or clustering?

- Classification
  - Because you know the correct classification for hundreds of images, you can build a model to classify the rest (supervised learning).
- Clustering

5. Basic Data Preparation

Knowledge Check 5.2: Outlier Detection



Which point is the outlier?

- The leftmost point
  - The leftmost point is well-aligned with the rest of the data, and its coordinates are close to others.
- The rightmost point
  - The rightmost point is well-aligned with the rest of the data, and its coordinates are close to others.
- The highest point
  - The highest point's vertical-axis value is far from the vertical-axis value of any of the other points.

Knowledge Check 5.3: Dealing with Outliers

Which of these is generally a good reason to remove an outlier from your data set?

- The outlier is an incorrectly-entered data, not real data.
  - If the data point isn't a true one, you should remove it from your data set.
- Outliers like this only happen occasionally.
  - If the data point is showing a real effect that occasionally occurs, it's usually important to account for that effect in your analysis.

6. Change Detection**Knowledge Check 6.1: Introduction to Change Detection**

Why are hypothesis tests often not sufficient for change detection?

- They don't really detect changes.
    - Hypothesis tests can detect changes.
- They often are slow to detect changes.
    - Hypothesis tests generally have high threshold levels, which makes them slow to detect changes.

Knowledge Check 6.2: CUSUM for Change Detection

In the CUSUM model, having a higher threshold T makes it…

- …detect changes faster, and less likely to falsely detect changes.
    - A higher threshold makes it slower to detect changes.
- …detect changes faster, and more likely to falsely detect changes.
    - A higher threshold makes it slower to detect both true and false changes.
- …detect changes slower, and less likely to falsely detect changes.
    - A higher threshold makes it slower to detect both true and false changes
- …detect changes slower, and more likely to falsely detect changes.
    - A higher threshold makes it lower to falsely detect changes.

Knowledge Check 6.3: Change-Detection Homework Followup

Analytics professionals are trusted to report the true findings of their models and analyses.

- True
    - So, it's important to always be honest about results.
- False
    - Most people don't have the training we do, so they trust us to tell them the truth.

7. Time Series Models

Knowledge Check 7.1: Introduction to Exponential Smoothing

In the exponential smoothing equation $S_t = \alpha x_t + (1-\alpha)S_{t-1}$ a value of closer to 1 is chosen if…

- There’s less randomness, so we’re more willing to trust the observation x_t
  - We put more weight on the observation x_t than the previous estimate S_{t-1}
- There’s more randomness, so we’re more willing to trust the previous estimate S_{t-1}
  - High puts more weight on the observation x_t than on the previous estimate S_{t-1}

Knowledge Check 7.2: Trends and Cyclic Effects

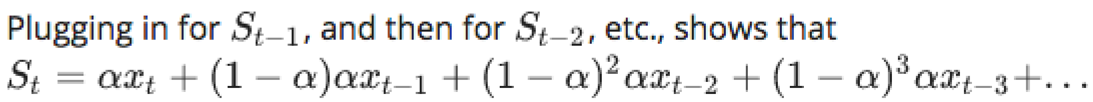
A multiplicative seasonality, like in the Holt-Winters method, means that the seasonal effect is…

- The same regardless of the baseline value.
- Proportional to the baseline value.
  - A multiplicative seasonality is larger when the baseline value is larger, because its effect is a multiple of the baseline

Knowledge Check 7.3: Exponential Smoothing: What Name Means

In the exponential smoothing equation $S_t = \alpha x_t + (1-\alpha)S_{t-1}$ only the current observation xt is considered in calculating the estimate S_t.

- True
- False

- 

  Plugging in for S_{t-1}, and then for S_{t-2}, etc., shows that

  $$S_t = \alpha x_t + (1-\alpha)\alpha x_{t-1} + (1-\alpha)^2 \alpha x_{t-2} + (1-\alpha)^3 \alpha x_{t-3} + \ldots$$

Knowledge Check 7.4: Forecasting

Is exponential smoothing better for short-term forecasting or long-term forecasting?

- Short-term
  - Exponential smoothing bases its forecast primarily on the most-recent data points. For forecasts of the longer-term future, there aren’t data points close to the time being forecasted
- Long-term

Knowledge Check 7.5: ARIMA

What does autoregression mean?

- The regression is done automatically
  - While software can automatically calculate regression coefficients for you, this isn’t what ‘auto’ means in "autoregression".
- Previous values of the thing being estimated are used to calculate the estimate
  - Its own previous values are used in the estimate
- The regression model is about cars
  - “Auto” has a different meaning here

Knowledge Check 7.6: GARCH

Why is GARCH different from ARIMA and exponential smoothing?

- GARCH uses time series data
  - GARCH, ARIMA, and exponential smoothing all use time series data.
- GARCH is autoregressive
  - GARCH and ARIMA are both autoregressive
- GARCH estimates variance
  - ARIMA and exponential smoothing both estimate the value of an attribute; GARCH estimates the variance

8. Basic Regression

Knowledge Check 8.1: Introduction to Regression

When would regression be used instead of a time series model?

- When there are other factors or predictors that affect the response
    - Regression helps show the relationships between factors and a response
- When only previous values of the response affect its current value
    - If only previous values of the response affect its current value, then a time series model is more appropriate

Knowledge Check 8.2: Maximum Likelihood and Information Criteria

If two models are approximately equally good, measures like AIC and BIC will favor the simpler model. Simpler models are often better because…

- Simple models are easier to explain and “sell” to managers and executives
    - True, but there are other reasons too
- The effects observed in simple models are easier for everyone, including analytics professionals, to understand
    - Usually true, but there are other reasons too
- Simple models are less likely to be over-fit to random effects
    - True, but there are other reasons too
- All of the above
    - Simpler models are less likely to be over-fit, easier to understand, and easier to explain

-Knowledge Check 8.3: Using Regression

Which of the following is not a common use of regression?

- Descriptive analytics: Understanding how the values of attributes relate to the value of a response.
    - Regression is often used for descriptive analytics
- Predictive analytics: Predicting the value of a response given the values of attributes.

    - Regression is often used for predictive analytics
- Prescriptive analytics: Determining the best course of action
    - Regression is often good for describing and predicting, but is not as helpful for suggesting a course of action

Knowledge Check 8.4: Causation vs Correlation

True or false: regression is a way to determine whether one thing causes another.

- True
- False
    - Regression can show relationships between observations, but it doesn't show whether one thing causes another

Knowledge Check 8.5: Transformation and Interactions

Suppose our regression model to estimate how tall a 2-year-old will be as an adult has the following coefficients:
0.56xFatherHeight + 0.51xMotherHeight – 0.02xFatherHeightxMotherHeight
The negative sign on the coefficient of FatherHeightxMotherHeight means:

- People with two taller-than-average parents won't be as tall as the individual effects of father's height and mother's height add up to
    - The negative coefficient for the interaction term brings down the overall estimate
- People with two taller-than-average parents will be taller than the sum of the individual effects of father's height and mother's height

Knowledge Check 8.6: Regression Output

True or false: A model with an R2 value lower than 0.8 is useless.

- True
- False

- Most real-life systems have so much uncertainty and so many factors that such high R2 values are very rare, and there’s often significant value in understanding even 20-30% of the variability

9. Advanced Data Preparation

Knowledge Check 9.1: Box-Cox Transformation

What does “heteroscedasticity” mean?

- The variance is different in different ranges of the data correct
    - Correct! And it’s not even such a hard word to say compared to ‘pneumonoultramicroscopicsilicovolcanoconiosis’!
- The variances of two samples of data are different from each other
    - Heteroscedasticity usually refers to just one data set.

Knowledge Check 9.2: De-Trending

You might want to de-trend data before…

- …using time-series data in a triple exponential smoothing (Holt-Winters) model
    - Triple exponential smoothing accounts for trend itself, so you don’t need to de-trend the data first
- …using time-series data in a regression model
    - Factor-based models like regression generally don’t account for time-based effects like trend.
- …using non-time-series data in a regression model
    - By definition, trend is only relevant for time-series data.

Knowledge Check 9.3: Introduction to Principal Component Analysis

Which of the following does principal component analysis (PCA) do?

- Transform data so there’s no correlation between dimensions.
    - True, but it does more than that.
- Rank the new dimensions in likely order of importance.
    - True, but it does more than that.
- Both of the above choices.
    - PCA can be a useful tool.

Knowledge Check 9.4: Using Principal Component Analysis

If you use principal component analysis (PCA) to transform your data and then you run a regression model on it, how can you interpret the regression coefficients in terms of the original attributes?

- The first coefficient corresponds to the first attribute in your original data set, the second coefficient corresponds to the second attribute, etc.
    - Each coefficient is for one of the principal components, not one of the original attributes.
- Each original attribute's implied regression coefficient is equal to a linear combination of the principal components' regression coefficients.
    - This is equivalent to using the inverse transformation.

10. Advanced Regression

Knowledge Check 10.1: Introduction to CART

1. True or false: In a regression tree, every leaf of the tree has a different regression model that might use different attributes, have different coefficients, etc.

- True
    - Each leaf's individual model is tailored to the subset of data points that follow all of the branches leading to the leaf.
- False
    - Each leaf's individual model is tailored to the subset of data points that follow all of the branches leading to the leaf. In some cases, it might mean that different leaves' models even use entirely different sets of attributes.

2. True or false: Tree-based approaches can be used for other models besides regression.

- True
    - For example, a classification tree might have a different SVM or KNN model at each leaf. It might even use SVM at some leaves and KNN at others (though that's probably rare).
- False
    - Tree-based approaches can be applied to a lot more than just regression.

Knowledge Check 10.2: Branching

A common rule of thumb is to stop branching if a leaf would contain less than 5% of the data points. Why not keep branching and allow models to find very close fits to each very small subset of data?

- Actually, that sounds like a great idea – we should keep branching and let models find very close fits to very small subsets of data!
    - No, we shouldn't. The models will experience overfitting – with too few data points, they'll fit to random patterns as well as real ones (see the Validation module).
- Fitting to very small subsets of data will cause overfitting.
    - With too few data points, the models will fit to random patterns as well as real ones.

- Fitting to very small subsets of data will make the tree have too many leaves.
  - With too few data points, the models will fit to random patterns as well as real ones (see the module on Validation).

Knowledge Check 10.3: Random Forests

True or False: When using a random forest model, it's easy to interpret how its results are determined.

- True
- False
  - Unlike a model like regression where we can show the result as a simple linear combination of each attribute times its regression coefficient, in a random forest model there are so many different trees used simultaneously that it's difficult to interpret exactly how any factor or factors affect the result.

Knowledge Check 10.4: Logistic Regression

a logistic regression model can be especially useful when the response…

- …is a probability (a number between zero and one).
  - Logistic regressions use a function that returns a value between zero and one, but they can be more broadly useful than this.
- …is binary (either zero or one).
  - True, but logistic regressions can also be useful for predicting values between zero and one.result.
- …can take any value.
  - Logistic regressions use a function that returns a value between zero and one.
- Both of the first two answers.
  - Logistic regressions can be useful for either situation.

Knowledge Check 10.5: Confusion Matrices

A model is built to determine whether data points belong to a category or not. A "true negative" result is:

- A data point that is in the category, but the model incorrectly says it isn't.
  - This is a false negative. 'True' and 'false' refer to whether the model is correct or not

- A data point that is not in the category, but the model incorrectly says it is.
    - This is a false positive. ‘True’ and ‘false’ refer to whether the model is correct or not, and ‘positive’ and ‘negative’ refer to whether the model says the point is in the category.
- A data point that is in the category, and the model correctly says it is.
    - This is a true positive. ‘Positive’ and ‘negative’ refer to whether the model says the point is in the category.
- A data point that is not in the category, and the model correctly says so.
    - True’ and ‘false’ refer to whether the model is correct or not, and ‘positive’ and ‘negative’ refer to whether the model says the point is in the category.
- A “Debbie Downer” (someone who often says negative things that bring down everyone’s mood).

Knowledge Check 10.6: Situationally-Driven Comparison

True or False: The most useful classification models are the ones that correctly classify the highest fraction of data points.

- True
- False
    - Sometimes the cost of a false positive is so high that it’s worth accepting more false negatives, or vice versa.

11. Variable Selection

Knowledge Check 11.1: Introduction to Variable Selection

Building simpler models with fewer factors helps avoid which problems?

- Overfitting
  - Using too many factors can sometimes lead to overfitting.
- Low prediction quality
  - It's possible that using fewer factors to avoid overfitting could slightly improve the prediction quality, but that's not one of the main effects.
- Bias in the most important factors.
  - Assuming the factors you'd drop aren't the most important ones, you'll still be stuck with the bias.
- Difficulty of interpretation
  - An overly-complex model can be hard to interpret, especially when factors are correlated with each other.

Knowledge Check 11.2: Models for Variable Selection

Which of these is a key difference between stepwise regression and lasso regression?

- Lasso regression requires the data to first be scaled.
  - If the data is not scaled, the coefficients can have artificially different orders of magnitude, which means they'll have unbalanced effects on the lasso constraint.
- Stepwise regression gives many models to choose from, while lasso gives just one.
  - At each step, the stepwise regression fits a different model. However, different lasso models can be found by varying T, and R has a function to automatically generate multiple lasso models.

Knowledge Check 11.3: Choosing a Variable Selection Model

When two predictors are highly correlated, which of the following statements is true?

- Lasso regression will usually have positive coefficients for both predictors.
    - Lasso will generally choose just one of them for the model. The choice is arbitrary, so we might end up having the worse choice as part of the model, with the better one left out.
- Ridge regression will usually have positive coefficients for both predictors.
    - Ridge regression will choose smaller positive coefficients for both models. By nature, it may underestimate the effect of the factors.

12. Design of Experiments

Knowledge Check 12.1: Introduction to Design of Experiments

If we're testing to see whether red cars sell for higher prices than blue cars, we need to account for the type and age of the cars in our data set. This is called

- Controlling
  - We need to control for the effects of type and age.
- Comparing
  - We are comparing red and blue cars, but there's a special term for accounting for the effects of other factors.
- Combining
  - We're not combining. There's a special term for accounting for the effects of other factors.

Knowledge Check 12.2: A/B Testing

In which of these scenarios would A/B testing not be a good model to use?

- Data can be collected quickly
  - Collecting data quickly enough is necessary for A/B testing to be successful.
- The collected data is not representative of the population we want an answer about.
  - If the data is not representative of the population, then A/B testing will not give a reliable answer.

Knowledge Check 12.3: Factorial Designs

In which of these situations is a factorial design more appropriate than A/B testing?

- Choosing between two different banner ad designs for orange juice.
  - With just two choices, A/B testing is all we need.
- Picking the best-tasting of four different brands of orange juice.

    - When we’re just comparing entire alternatives, A/B testing (or a more general form – that tests more than two alternatives) is all we need.
- Finding the best combination of factors in orange juice to maximize sales.
    - Factorial design helps us find the impact of each different factor in combination with others, as opposed to A/B testing that just compares entire alternatives.

Knowledge Check 12.4: Multi-Armed Bandits

Which of these is a way that multi-armed bandit models deal with balancing exploration and exploitation?

- As we get more sure of the best answer, we’re more likely to choose to use it.
-
    - Multi-armed bandit models use the best answer (exploitation) the more they’re sure it’s best. If the model is less sure what’s best, it’s more likely to concentrate on trying many options (exploration).
- If we’re unsure of the best answer, we should pick one and stick with it.
    - The less sure a multi-armed bandit model is of the best answer, the more likely it is to try lots of possibilities (exploration) rather than focusing on one good answer (exploitation).
- As we get more sure of the best answer, we’re more likely to try many different ones just to make sure.

13. Probability-Based Models

Knowledge Check 13.2: Bernoulli, Binomial and Geometric Distributions

1. A nationwide lawn care company wants to estimate the number of days each month that the temperature is above 50 degrees Fahrenheit. Why is a binomial distribution (n = number of days in the month, p = probability of the temperature being above 50) not a good model for them to use?

- The binomial distribution models the number of “yes” answers out of some number of observations.
  - it’s a correct statement, but it doesn’t explain why a binomial distribution is inappropriate in this case. Our case does ask for the number of ‘yes’ answers (days above 50 degrees) out of some number of observations (the number of days in the month).
- The results aren’t independent – days above 50 degrees are more likely to be clumped together in the summer, for example.
  - Another way of describing the same effect is that the value of p will differ from month to month, and the binomial distribution requires it to be the same)

2. Let p be the probability of a successful sales call. (For this question, assume the probability is the same for each sales call, and the results for different sales calls are independent.) Which of the following statements is true?

- According to the geometric distribution, the probability of having 5 successful sales calls before the first unsuccessful call is $(1-p)^5p$
- According to the geometric distribution, the probability of having 5 successful sales calls before the first unsuccessful call is $p^5(1-p)$
  - The probability of having 5 successful sales calls is p^5, and then the subsequent unsuccessful call has probability $(1-p)$

Knowledge Check 13.3: Poisson, Exponential and Weibull Distributions

1. If the time between customer arrivals to a restaurant at lunchtime fits the exponential distribution, then which of the following is true?

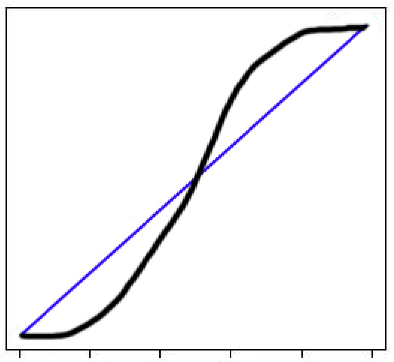
- The number of arrivals per unit time follows the Weibull distribution.
  - The exponential distribution is the same as the Weibull distribution with k=1, but that doesn’t tell us the distribution of the number of arrivals per unit time.
- The number of arrivals per unit time follows the Poisson distribution.

    - If the interarrival time is exponentially distributed, then the number of arrivals per unit time follows the Poisson distribution – and if the number of arrivals per unit time is Poisson, then the interarrival times are exponentially distributed.

2. What is the difference between the interpretations of the geometric and Weibull distributions?

- The geometric distribution models how many tries it takes for something to happen, while the Weibull distribution models how long it takes.
    - True. As a consequence, the geometric distribution takes only integer values, while the Weibull distribution is continuous.
- The Weibull distribution models how many tries it takes for something to happen, while the geometric distribution models how long it takes.
    - No, it's the opposite. As a consequence, the geometric distribution takes only integer values, while the Weibull distribution is continuous.

Knowledge Check 13.4: Q-Q Plots



The figure above is a Q-Q plot of a data set compared to the normal distribution. What does the plot show?

- The data has more points at each end (the "tails") than would fit the normal distribution.
    - This is called a 'heavy-tailed' distribution.
- The data set is a good fit to the normal distribution.
    - The plot does not closely follow the line, so the data is not a good fit to the distribution.

- The data has fewer points at each end (the “tails”) than would fit the normal distribution.
    - The data set shows more points at the tails than the normal distribution would have.

Knowledge Check 13.5: Queuing

Which of these is a queuing model not appropriate for?

- Determining the average wait time on a customer service hotline.
    - This can be modeled as a queue. For example, customer calls arrive and wait for the next available customer service agent, who answers their call.
- Estimating the length of the checkout lines at a grocery store.
    - This can be modeled as a queue. For example, customers arrive at the checkout area, and choose a checkout line to wait in (each line might be a separate queue).
- Predicting the number of customers who will come to a restaurant tomorrow.
    - This question doesn’t involve waiting in line for service. A queuing model is not appropriate.

Knowledge Check 13.6: Simulation Basics

1. Why should a stochastic simulation be run many times?

- To verify that the same thing happens each time.
    - The point of a stochastic simulation is to study the effect of variability and randomness – we don’t want every run to be exactly the same, so we can study the overall system performance in different situations that could arise.
- One random outcome might not be representative of system performance in the range of different situations that could arise.
    - A stochastic simulation is meant to show the performance of a system over a range of random events that could happen.

2. Why is it important to validate a simulation by comparing it to real data as much as possible?

- If the simulation isn't a good reflection of reality, then any insights we gain from studying the simulation might not be applicable in reality.
    - This is such an important point about simulation that I wanted to make sure you all clicked on it. I've overseen a lot of analytics projects that included simulation, and my experience is that it's easy to rely too much on simulated insights that might not be true in reality. It's critical to make sure that the simulation is a good-enough model of reality that insights from the simulation can effectively be transferred to reality

Knowledge Check 13.7: Prescriptive Simulation

How can simulation be used in a prescriptive analytics way, for example to determine the right number of voting machines and poll workers at an election location?

- Vary parameters of interest (like the number of voting machines and the number of poll workers), compare the simulated system performance, and select the setup with the best results (for example, the best balance between low waiting times to vote and low cost of machines and workers).
    - This is a correct answer, but it's not the only one. Most simulation software now has automated functions to do this for you.
- Use the automated optimization function in simulation software (for example, OptQuest in Arena) to find parameter values that give good results.
    - This is a correct answer, but it's not the only one. You can also do this sort of thing by and..
- Both of the answers above.
    - Correct! You can do it by hand, or using the built-in functions.

Knowledge Check 13.8: Markov Chains

In analytics, the term "memoryless" means

- The next state of a process doesn't depend on its current state or any of its previous states.
    - It's true that the next state of a 'memoryless' process doesn't depend on previous states, but it does depend on the current state.

- The next state of a process doesn’t depend on any of its previous state, but does depend on the current state.
    - The next state of a ‘memoryless’ process doesn’t depend on previous states, but it does depend on the current state
- I don’t remember. Does that mean I’m memoryless?

14. Missing Data

Knowledge Check 14.1: Introduction to Missing Data

Which of these is a common reason that data sets are missing values?

- A person accidentally typed in the wrong value.
- A person did not want to reveal the true value.
- An automated system did not work correctly to record the value.
- All of the above.

Knowledge Check 14.2: Methods That Do Not Require Imputation

Which of the following statements is not a situation where missing data can be biased?

- Due to a programming flaw, a security camera discards every 7th frame of video it takes.
    - There’s no pattern to what security-related data is kept or discarded.
- People might be less willing to express certain political views to survey-takers, or less willing to report certain incomes.
    - This is an example of how certain values could be more likely than others to be missing.
- GPS devices might be more likely to lose service in some areas than in others.
    - This is an example of how certain values could be more likely than others to be missing.

Knowledge Check 14.3: Imputation Methods

Which statements are true about data imputation?

- Imputing more than 5% of values is usually not recommended.
- Imputation of more than one factor value for a data point is also possible.
- Both answers above are correct.
    - True! It’s usually not recommended to impute more than 5% of values, and advanced methods like multivariate imputation by chained equations (MICE) can impute multiple factor values together

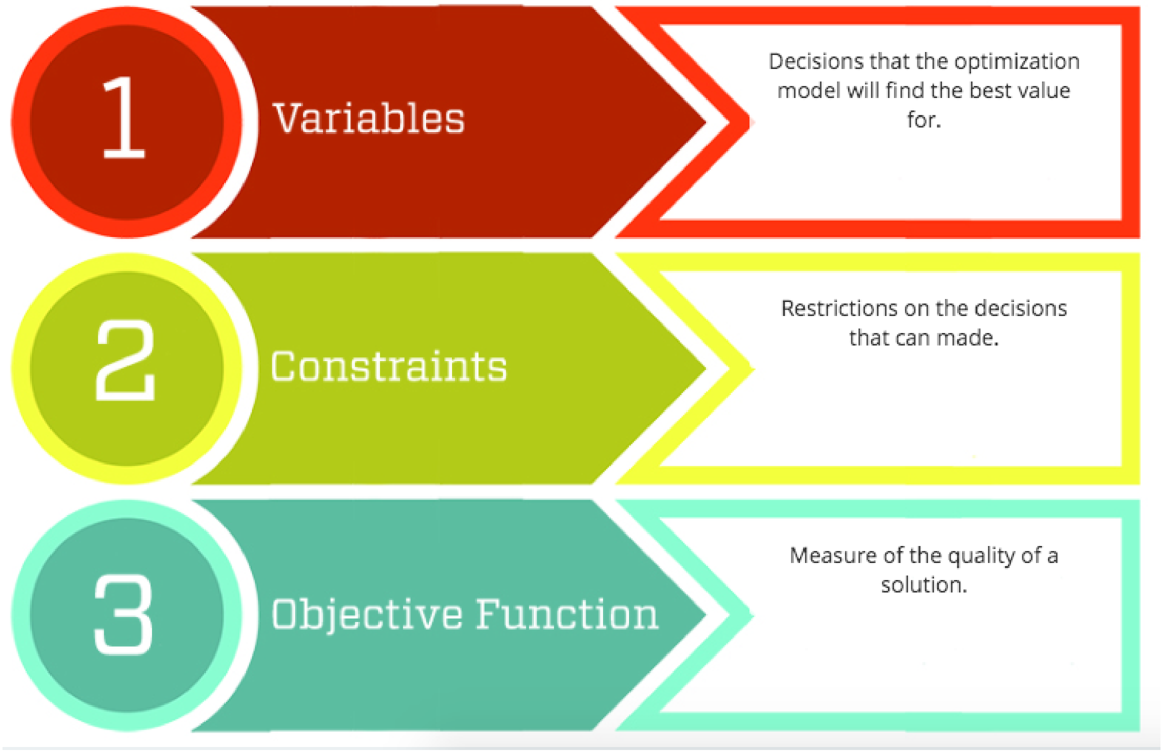
15. Optimization

Knowledge Check 15.1: Introduction to Optimization

Which of the following is true?

- Statistical software can both build and solve regression models. Optimization software only solves models; human experts are required to build optimization models.
- Optimization software can both build and solve models. Regression software only solves models; human experts are required to build regression models.

Knowledge Check 15.2: Elements of Optimization Models


1
Variables
Decisions that the optimization model will find the best value for.
2
Constraints
Restrictions on the decisions that can made.
3
Objective Function
Measure of the quality of a solution.

Knowledge Check 15.4: Modeling With Binary Variables

Let y_i and y_j both be binary variables, and let x_i be a continuous variable. Which expression is equivalent to the constraint "If $y_i = 1$, then y_j must also be 1?

- $y_i + y_j \leq 1$
  - ○ No, this constraint says y_i and y_j can't both be 1.
- $y_j \geq y_i$
  - ○ If $y_i = 1$, this constraint becomes $y_j \geq 1$. Since y_j must be either 0 or 1, the only possibility is for y_j to be 1.
- $x_i \leq By_i$

    - y_j is not even part of this constraint

Knowledge Check 15.5: Optimization for Statistical Models

Which of these statistical models does not have an underlying optimization model to find the best fit?

- Linear regression
- Logistic regression
- Lasso regression
- Exponential smoothing
- k-means clustering
- None of the above
    - All of these statistical models have underlying optimization models.

Knowledge Check 15.6: Classification of Optimization Models

True or false: Requiring some variables in a linear program to take integer values can make it take a lot longer to solve.

- True
    - Adding integer variables moves the model from a linear program, which usually solves very quickly, to an integer program, which sometimes takes a long time to solve.
- False

Knowledge Check 15.7: Stochastic Optimization

True or false: Optimization models implicitly assume that we know all of the values of the input data exactly.

- True
    - Optimization models treat all of the data as known exactly.
- False

Knowledge Check 15.8: Basic Optimization Algorithms

The two main steps of most optimization algorithms are:

- Find the most important remaining variable, and assign its value to be as large as possible.
    - No, giving important variables large values is not always a good idea.
- Find a good direction to move from the current solution, and determine how far to go in that direction.
    - Many optimization algorithms follow the pattern of finding an improving direction and a step size, make the move, and repeat.
- Pick a set of variables to be zero, and find the best values of the remaining variables.
    - The simplex algorithm for solving linear programs can be thought of this way, but these steps aren't used in most optimization algorithms.

16. Advanced Models

Knowledge Check 16.1: Non-Parametric Methods

Nonparametric tests are useful when…

- we don't know much about the form of the underlying distribution the data comes from, or it doesn't fit a nice distribution.
    - Because we don't have a good distribution to fit parameters to, a nonparametric test is useful.
- it's important to have information about the median.
    - Many nonparametric tests focus on the median, and they can be used even when we do not know the form of the underlying distribution
- we don't have much data.
    - This one is a little tricky! By focusing on the median, nonparametric tests make it less important whether a small data set includes the right

distribution and range of results. All a nonparametric test needs is enough data to figure out approximately where the middle value is.

- we want to know whether the means of two distributions are similar.
  - Nonparametric tests usually focus on the median, not the mean.

Knowledge Check 16.2: Bayesian Modeling

How can Bayesian models incorporate expert opinion when there's not as much data to analyze as we'd like to have?

- In Bayes' theorem $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$, both A and B are parameters that can be adjusted by an expert.
  - A and B are things that happen or are observed – for example, A might be that someone has a disease, and B might be whether that person tested positive for the disease.
- Expert opinion can be used to define the initial distribution of P(A), and observed data about B can be used with Bayes' theorem to obtain a revised opinion P(A|B).
  - The initial distribution assumed for P(A) is called the 'prior distribution' and the revised distribution P(A|B) is called the 'posterior distribution'

Knowledge Check 16.3: Communities in Graphs

(Advanced question – take a few minutes to think) Suppose we have a graph where all edge weights are equal to 1. In the video, we saw how to split a graph up into highly-interconnected communities. Now, instead we want to split the nodes into large groups that have very few connections between them (for example, if a marketer wants to find sets of people in a social network who probably have very different sets of friends). How might you do that?

- It's not possible using what we've covered in the video.
- Change the Louvain algorithm to minimize modularity instead of maximizing it.

    - Good try! But having the Louvain algorithm minimize would just result in no changes being made: each node will remain its own community, rather than grouping nodes that aren’t connected.
- Change the graph: for every pair of nodes i and j, if there’s an edge between i and j then remove it; and if there’s not an edge between i and j, then add it. Then run the Louvain algorithm on the new graph.
    - Good thinking! This will find highly-interconnected communities in the new graph, which are equivalent to very-low-connectivity communities in the original graph. [Quick jargon break: just like a set of nodes with edges between each pair is called a ‘clique’, a set of nodes without any edges between them is called an ‘independent set’.]

Knowledge Check 16.4: Neural Networks and Deep Learning

Deep learning is one of the current best approaches for which of these?

- Image recognition
    - Deep learning is currently one of the best approaches for recognizing images, speech, writing, and language.
- Splitting graphs into highly-connected communities
    - An approach like the Louvain algorithm is more appropriate for splitting graphs into highly-connected communities.
- Demand forecasting
    - At this time, other forecasting models we’ve seen in this course usually do better than deep learning.

Knowledge Check 16.5: Competitive Models

Which of the following is a situation where competitive decision-making (e.g., a game theoretic model) would be appropriate?

- A company wants to optimize its production levels, based on production cost, price, and demand. The company already has estimated a function to give predicted selling price and demand as a function of the number of units produced.

  - The price and demand are functions of production, so the only decision to be made is the company’s decision of how much to produce.
- A company wants to optimize its production levels, based on production cost, price, and demand. The company already has estimated a function to give predicted selling price and demand as a function of the number of units produced, and the number of units its competitor produces.
  - The price and demand are functions of the competitor’s decision, so both company’s decisions must be accounted for.

For the weeks between the last week and this one, I could not find any knowledge checks for any of the Case Studies (Weeks 12 to 14). Please add them if you have some for these sections.

Lesson 21.1 (C): Many Analysts, One Dataset

In analytics modeling, we need to…

- …try to avoid our biases affecting our models.
  - Yes, can be important – even if the question is simple (e.g., will people like a blue banner ad better than a white one) we need to make sure our own thoughts (e.g., I think white backgrounds look nice) aren’t biasing our models.
- …report our results honestly.
  - People are trusting us as professionals.
- …be open to other people’s models, even if they’re different from ours.
  - Another approach might turn out to be better. But you certainly should critique someone’s model if it’s incorrect or used inappropriately! There might be more than one right answer, but there can also be lot of wrong ones, and part of our job is to differentiate between them.
- …develop our own intuition and artistry.
  - Modeling is an art!