**Knowledge Checks**

**The Big Picture**

**m1q1**

1.0/1.0 point (graded)

This course's approach to instructing data science is through:

Explaining how we would go about solving data driven problems, then having you mimic our work.

Arming you with the know-how needed solve data driven problems, then having you attempt to do so.  
correct

Having you watch videos over and over until you're able to understand how to solve data driven problems.

None of the above.

**m1q2**

1.0/1.0 point (graded)

Which statement makes the most sense about data analysis?

Evaluate your data before you transform it, so that you only spend time transforming worthy data.

The most critical step the modeling step; everything else is extra.

You can solve just about any problem with data analysis, as long as you use machine learning models.

Special care ought be dedicated to collecting data, so that you have enough of it to do effective analysis.  
correct

You have to be a professional programmer before attempting to do any sort of data analysis.

**m1q3**

1.0/1.0 point (graded)

If you want to engage the data analysis process, the best place to start is by...

Making your data more reliable by getting rid of bad entries.

Exploring your data with visualizations, to see if analysis is even possible.

Collecting data from various sources.

Evaluating how likely data analysis is to work with your data.

Having a question in mind to drive your data collection process.  
correct

**m1q4**

1.0/1.0 point (graded)

What is the difference between supervised and unsupervised learning?

With unsupervised learning, it is up to you if you'd like to observe the computer train against data or not.

Difference? There is no difference.

Unsupervised learning attempts to extract patterns; supervised learning tries to fit rules and equations.  
correct

With supervised learning, it is actually possible to observe the computer train against data samples.

**m1q5**

1.0/1.0 point (graded)

Select the three best applications of machine learning:

Getting computers to solve data driven problems.

Using computers to solve tasks that involve fine-tuning many parameters.

Having the computer automatically grade thousands of multiple choice questions.

When you have many observations to train a computer against.

correct

**m1q6**

1.0/1.0 point (graded)

The small credit union / bank you've worked at for the past ten years is having cash-flow issues. Upon further inspection, it turns out most of their clients are defaulting on loans! In trying to rectify the business, you've been assigned the task of studying all client data, and then suggesting a method to identify which prospective borrowers are likely to default on their loans.

What type of machine learning would you engage to solve this problem?

Dimensionality Reduction

Classification  
correct

Clustering

Reinforcement Learning

Regression

**m1q7**

1.0/1.0 point (graded)

After doing a great job of solving the cash flow issue by denying loans to bad clients, your boss issued a client satisfaction survey to ensure that those who remained loyal customers were happy with all the new process changes. Apparently, he was really motivated while creating the survey because he came up with thirty questions. Even more amazingly, almost all the clients responded! Now your weekend plans are ruined because you have to go through thousands of survey responses to summarize how the clients are feeling about the credit union.

What type of machine learning would you engage to solve this problem?

Dimensionality Reduction  
correct

Classification

Clustering

Reinforcement Learning

Regression

**m1q8**

1.0/1.0 point (graded)

Which of the following is an example of regression?

Categorizing polar bear and grizzly bear footprints.

Calculating the relationship between a black hole's mass and its Hawking radiation.  
correct

Grouping meals together by similar ingredient.

Taking a list of people's like / dislike responses to 200 movies and creating a list of 5 movie genres they either like or dislike.

Learning to walk by minimizing a penalty score metric.

**Data & Features**

**m2q1**

1.0/1.0 point (graded)

If you were looking at a dataset of 'Companies', which **two** of the following features would you expect to be stored as **textual** data types?

Employee Count

Street

Days in Business

Zip Code

CEO

correct

**m2q2**

1.0/1.0 point (graded)

Which of the following is an example of an *ordinal* feature?

The various colors of bed sheets

The point score a student receives on their final exam

The grade-level a student belongs to (e.g. 1st grade, 2nd grade, ...)  
correct

The school name a student belongs to

The student's telephone number

**m2q3**

1.0/1.0 point (graded)

Which of the following is most suited to be a continuous feature?

Airplane fuel gauge reading  
correct

Number of keys on a keyboard

Restaurant rating

Number of pens in a box

Computer manufacturer name

SaveSave Your Answer Show Answer

SubmitSubmit Your Answer

You have used 1 of 2 attempts

**m2q4**

1.0/1.0 point (graded)

The best features are...

Your raw data

Those that are highly correlated

Those with the most variance

Those with the least variance

Those pertaining to the task you're trying to solve  
correct

**m2q5**

1.0/1.0 point (graded)

Given a dataset with many features, consisting of patient medical records, let's say you want to trim it down to only those features that are related to patient allergies. Which **one** of the following features is **least likely** to be statistically relevant to a person's allergies?

Patient Drug Prescriptions

Patient Fingernail Length  
correct

Patient DNA

Patient Environment

**m2q6**

1.0/1.0 point (graded)

When building out your dataset initially, what are the **three** things should you focus on the most?

Collecting more features than samples, so that each observation is documented in great detail.

Collecting more samples than features, so that the mathematics required for machine learning works out well.

Collecting features, even if independently they don't do a great job at answering your dataset's question.

Ensure you only collect features that directly answer your question without ambiguity.

Let your intuition about the question, and expertise in the domain of your issue, drive you to choosing right features.

correct

**m2q7**

1.0/1.0 point (graded)

What is the difference between df.colA and df[['colA']]?

There is no difference, they are essentially equal.

One returns a series, the other returns a dataframe.  
correct

One executes much faster than the other.

The second one causes a syntax error.

**m2q8**

1.0/1.0 point (graded)

>>> import pandas as pd

>>> animals = pd.read\_csv('animals.csv', sep='\t')

>>> animals.columns

Given the above, fictional dataset, which command would you use to select the: *family*, *species*, and *population*columns of the first five rows?

animals.iloc[['family', 'species', 'population'], 0:6] 

animals.iloc['family', 'species', 'population', 0:4] 

animals.loc[0:4, ['family', 'species', 'population']]   
correct

animals.loc[1:5, ['fami ly', 'species', 'population']] 

animals.loc[0:5, 'family', 'species', 'population'] 

**m2q9**

1.0/1.0 point (graded)

features = [

"The enchanted forest beamed with magic once the prince was born.",

"Jinto's life changed forever when his planet surrendered without firing a single shot."

]

Which technique would you use to encode the above features?

Encode using CountVectorizer  
correct

Encode each word using increasing, ordered integers

Encode each word using increasing, ordered integers

Use a separate boolean column to encode each word

**m2q10**

1.0/1.0 point (graded)

Ordinal and Nominal features should be encoded using:

.astype()   
correct

.get\_dummies() 

.fit\_transform() 

.reshape() 

**m2q11**

1.0/1.0 point (graded)

A dataset of **oil wells** was made available to you in the form of an html table, which you've loaded into a Pandas dataset. While inspecting the dataset, you notice a *would-be* numeric features called 'depth' actually has a handful of entries that show up as '**?**' rather than as numbers.

What is the *best* course of action you should take to correct this?

No action is needed, as this was the desired result

Run a command to replace all the **?**'s with **0**'s, such as df.depth[df.depth=='?'] = 0

Update your .read\_html() method so that **?** is included as a **na\_value**, causing them to be converted to NANs

Run df.depth = pd.to\_numeric(df.depth, errors='coerce') to convert the column to a numeric type and coerce non-numeric values to NANs  
correct

**Exploring Data**

### m3q1

1.0/1.0 point (graded)

Suppose you have a dataset that listed burritos from various mom and pop's restaurants. Each sampling has ordinal features, which range from 1-10, and rank observations such as tortilla quality, volume, uniformity, meat quality, quantity, temperature, etc.

If you wanted to know how many burritos have a meat quality ranking of 6-8, which visualization method would you make use of?

A histogram  
correct

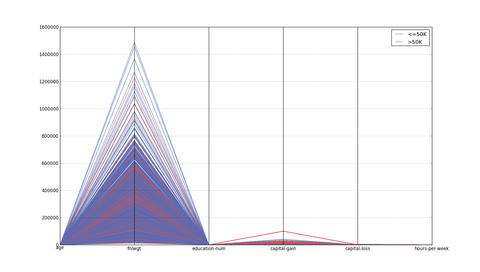
A 2d scatter plot

A 3d scatter plot

An imageshow

### m3q2

1.0/1.0 point (graded)



The above visualization is an example of...

A parallel coordinates plot, used effectively

An andrews plot, demonstrating colored grouping

An imshow rendering of a linear / line based correlation matrix

A parallel coordinates plot, suffering from feature scaling issues  
correct

### m3q3

1.0/1.0 point (graded)

You have the following questionnaire:

1. Rate BBC's "Planet Earth" television series (1-10)
2. Rate Discovery's "Man vs. Wild" television series (1-10)
3. Rate Paramount's "Star Trek Voyager" television series (1-10)

And a correlation matrix:

Q1 Q2 Q3

Q1 1.000000 0.707568 0.014746

Q2 0.707568 1.000000 -0.039130

Q3 0.014746 -0.039130 1.000000

Given the above questionnaire and correlation matrix generated from dataset of 1000 responses, what does the **0.707568** value tell you?

Planet Earth is, on average, rated higher than both Man vs. Wild and Star Trek Voyager

Star Trek Voyager is the best Star Trek series, since more people like it than the other two shows

People who like watching Planet Earth tend to enjoy watching Man vs. Wild  
correct

There isn't much of a correlation between any of the shows, since 1.00000's only appear in the diagonals

**Transforming Data**

**m4q1**

1.0/1.0 point (graded)

Please complete the sentence so that it makes the most sense:

Principal component analysis...

Requires you have labeled features to use as a metric for determining which features are most important

Is a dimensionality reduction technique that builds a simpler, non-linear projection or 'shadow' of your dataset

Asserts you have more features than samples so you can avoid the curse of dimensionality and the matrix math works out

Ensures each newly computed feature is orthogonal to all previously computed ones, minimizing overlaps  
correct

**m4q2**

1.0/1.0 point (graded)

Which of these statements is wrong?

PCA can be used to discover the underlying features being assessed by a dataset

The results of PCA depend on the scaling of your data, so having a feature with units of 'light-years' and another feature with units of 'GHz' may be disastrous

When applied to non-linear data, PCA generally isn't as effective as when applied to linear data

Since PCA is sensitive to feature scaling, if you have a feature that is a linear transformation of the other, e.g. feature2 = 10 \* feature1, then both features will be ignored  
correct

**m4q3**

1.0/1.0 point (graded)

Which of the following explanations of isomap is true?

Isomap can be used as a powerful noise removal tool, since a smooth manifold is created by "short circuiting" the nearest neighbor map when calculating distances

Isomap is usually faster than PCA because it's quicker to compute a nearest neighbor map than to do matrix decomposition

A one sentence summary of isomap's implementation is that at its core, it is essentially a node distance map that has been fed into a special type of PCA  
correct

Isomap will not function without a completely accurate distance metric, since it needs to know the precise distance to every single sample, including distant ones

**m4q4**

1.0/1.0 point (graded)

Isomap is most beneficial...

When your data lacks an inherent manifold

When a non-linear, geometric structure is expressed in your data  
correct

When you are uncertain how many samples are needed to capture the underlying nature of your data

When your high dimensionality data has a hidden, linear relationship expressed within it

**m4q5**

1.0/1.0 point (graded)

Given the following options:

1. If you encounter errors in your data, don't let anyone know and try to hide them by deleting the affected rows, or by cooking your data.
2. While gathering data, identify issues that might cause inconsistencies, and capture additional features that'll help you rectify them.
3. Retrospectively adjust your data to account for discovered problems.

Order these options from most desirable to least desirable:

A, B, C

A, C, B

B, A, C

B, C, A  
correct

C, A, B

C, B, A

**Data Modelling**

### m5q1

1.0/1.0 point (graded)

Only one of following statements is true. Which one is it?

It's possible for samples from two different clusters to be more similar to one another than their intra-cluster neighbors, if the two clusters are large and located near one another  
correct

Real world data typically comes labeled

Unsupervised clustering aims to group your samples based on their labels

Centroids are records that live in your dataset and share the same feature space so that a meaningful distance can be calculated between them and your samples

### m5q2

1.0/1.0 point (graded)

Only one of following statements is true. Which one is it?

It's possible for a sample to be assigned to two clusters; but only if its equidistant from either cluster.

The K-Means algorithm scans your dataset to detect clusters using an iterative assignment / update cycle. The algorithm returns the number of clusters found, as well as their centroid position.

As a clustering algorithm, K-Means is really only useful for grouping your samples

K-Means assumes your features are either length normalized, or that their length encodes a specific meaning.  
correct

### m5q3

1.0/1.0 point (graded)

One of the following statements is actually a lie. Select the statement that is inaccurate from the list below:

Overfitting means your machine learning algorithm is performing at 100% and has been over trained  
correct

In supervised learning, you provide the algorithm the correct answers while training it

If you split off too many samples for testing, your training is going to suffer as a consequence

There shouldn't be an overlap in your training and testing dataset, because your algorithm already has the answers to the training data

### m5q4

1.0/1.0 point (graded)

What of these demonstrates the proper order of operations?

Load Data, Encode Data / Wrangle Data, KMeans the Data, Split Data

Load Data, Encode Data / Wrangle Data, Split Data, Fit PCA with Training Data  
correct

Load Data, Encode Data / Wrangle Data, PCA the Data, Split Data

Load Data, Encode Data / Wrangle Data, Split Data, Fit Isomap with Testing Data

Load Data, Encode Data / Wrangle Data, Isomap the Data, Split Data

### m5q5

1.0/1.0 point (graded)

Overfitting is **best** described as:

Your machine learning model not generalizing well against new data  
correct

What happens when you don't split your data too much

Training your machine learning algorithms until they have a high level of accuracy

Working out more than necessary

### m5q6

1.0/1.0 point (graded)

Classification is the process of...

Looking for groups of samples based only on their features

Labeling samples depending based on their neighbors

Identifying the group membership of samples  
correct

Grouping similar samples, and then assigning a label or class to them

### m5q7

1.0/1.0 point (graded)

The main similarity between K-Neighbors and K-Means are...

They both use distance functions to tackle the problem of group assignment  
correct

They both have the same, non-linear, decision boundary

They both have a K in their names

They both are classification algorithms that aim to assign a label to your samples

### m5q8

1.0/1.0 point (graded)

Which of the following statements makes the most sense?

Linear regression helps you answer categorical questions

Linear regression can work with categorical features as inputs  
correct

Linear regression uses the same calculation that PCA does to get the shortest distance to a line / hyperplane

Linear regression allows you to extrapolate data more accurately than it allows you to interpolate it

### m5q9

1.0/1.0 point (graded)

The difference between the actual, observed y-value of your sample and the predicted y-value from the linear regression line is called?

A standard deviation

An error  
correct

A slope

Δy (Delta y)

A weight

**Data Modelling II**

### m6q1

0.0/1.0 point (graded)

The purpose of the kernel function in support vector machines is to...

Intelligently change your feature space from a linear one to polynomial one.

Increase the dimensionality of your data set.

Compute how similar two samples are.  
correct

Decrease the dimensionality of your data set, making it simpler to linearly separate your higher-dimensionality data

### m6q2

1.0/1.0 point (graded)

SciKit-Learn's SVC class takes in many parameters. Three of the parameters that contribute the most to your overall decision boundary shape are gamma, C, and the kernel. Assume you're using the rbf kernel and you like the overall shape of your decision boundary. However you notice it has 'pockets' (e.g. little bubbles) that surround one class of your samples. And those pockets are simply too large.

How would you alter your parameters fix this?

Change rbf kernel to linear kernel, so that no bubbles form anymore

Reduce the gamma coefficient to get rid of the bubbles  
correct

Increase C to get rid of the bubbles

Increase the gamma coefficient to get rid of the bubbles

**Hint:** You might want to check out Stanford's SVM-JS Demo, linked in the Dive Deeper section before answering this to visualize the effect of altering SVC parameters on your decision boundary.

### m6q3

1.0/1.0 point (graded)

**SKLearn's** Implementation of SVC...

Is structurally similar to the computation of linear regression

Is a purely analytical operation, so will always produce the same output given the same input features {{ that isn't what we stated in the reading.. }

Is a non-linear operation

Must be seeded in order to be truly deterministic  
correct

Is only effective in low dimensional spaces

### m6q4

1.0/1.0 point (graded)

Please mark the appropriate response below:

Decision trees are able to do non-linear classification because the tree branches don't use linear decision surfaces

Each node of the tree represents the outcome of decision, such as age > 18, leading to an eventual classification

The decision tree algorithm is able to build a tree based on the intrinsic geometry of your dataset

Decision trees do not use kernels, but are capable of doing non-linear classification by segmenting your feature-space.  
correct

### m6q5

1.0/1.0 point (graded)

Which of the following scenarios would you use decision trees to solve?

You want to build a model for your interns to use, for segmenting customers based on their reply to survey questions.  
correct

You want to discover which customers are more similar, for the purposes of targeted advertising.

Given a dataset of star masses and locations, you're interested in calculating the position of a galactic black hole.

Given pressure, temperature, and height above sea-level samples, you're interested in modeling the amount of oxygen present in the air.

### m6q6

1.0/1.0 point (graded)

How would you explain out-of-bag samples to someone who's studied decision trees but not yet random forest?

The out-of-bag samples are those samples withheld from the forest ensemble while training.

The out-of-bag samples are the bootstrapped samples used for training your decision trees

The out-of-bag samples are the bootstrapped samples used for training and evaluating the accuracy score of your random forest

The out-of-bag samples are those training samples withheld from a particular decision tree while training  
correct

The out-of-bag samples are those testing samples withheld from the ensemble so they can be used to test its accuracy

### m6q7

1.0/1.0 point (graded)

After telling your cousins you're taking a data science course that deals with machine learning, they get super excited and put together a fictitious dataset for you to run classification on to prove you know what you're doing. You decide to model their dataset using random forest, since that's the chapter you just studied. What's unsettling is that they didn't give you a lot of samples, but the samples they did provide have a **lot** of features. Due to this, you fear your RF model might be overfit, and there are many instances where long, erroneous-looking branches shoot out of the tree in order to correctly classify a single sample! Until now, you had thought the mere use of random forests was enough to inhibit overfitting; but it looks like that's not enough. Your cousins are coming over soon and can't wait to see the results of your modeling, and you don't want to let them down!

What parameter might you **increase** to stop the overfitting?

min\_samples\_split  
correct

criterion

max\_depth

max\_leaf\_nodes

n\_estimators

**Data Modelling II**

**m7q1**

1.0/1.0 point (graded)

After being hired by Robert Downey Sr's start-up "SpaceZ", you've been placed in charge of the artificial intelligence that controls the HyperExplorer Pod. This vessel will travel at very fast velocities through the asteroid belt to explore the long-term effects of low-gravity and space-travel on the humans, animals, and plants onboard. It's important the craft is able to maneuver around asteroids safely, but gas clouds aren't as important to avoid (we know gas clouds don't just hang out in space, but work with us here...) You have around 90-thousand data samples of various space instrument readings stored as features, all of which were collected from earlier satellite launches. The first item of business you need to take care of is identifying the type of objects the vessel encounters, such as comets, asteroids, aliens, gas clouds, and solar flares.

Which of the following algorithms makes the most sense to use to do this?

Decision Tree Regression

Linear Regression

Randomized PCA

K-Means

SVC  
correct

**m7q2**

1.0/1.0 point (graded)

Given the column definitions: Alligator, Crocodile, and Water Snake, and the following confusion matrix: Which target is your model doing a really bad job of correctly predicting, and what class does it actually think that target is?  
| 3 | 0 | 1 |  
| 4 | 1 | 1 |  
| 0 | 2 | 3 |

Crocodile, Alligator  
correct

Alligator, Crocodile

Crocodile, Water Snake

Water Snake, Alligator

Alligator, Water Snake

**m7q3**

1.0/1.0 point (graded)

If all you cared about was how *fast* an algorithm could come up with a decent approximation of how much wool a sheep produces per year given the altitude above sea-level it lives, how much it eats per day, and the amount of day-light hours it receives each month, and the existence (or not) of a few genetic markers...

Which of the following algorithms are you most likely to use?

Logistic Regression

Linear Regression  
correct

Boosted Decision Trees Regression

Isomap

Neural Networks

**m7q4**

1.0/1.0 point (graded)

If a rocket ship's navigation computer needs to classify asteroids to avoid them...

True positives are the most harmful, because an android will certainly strike the ship.

True negatives are the most harmful, because the model failed to classify the asteroid, causing it to strike the ship.

False positives are the most harmful, because the model got confused resulting in the ship getting struck.

False negatives are the most harmful, because the model didn't tell the ship to dodge the asteroid.  
correct

**m7q5**

1.0/1.0 point (graded)

Here are a couple of definitions; the name of these techniques have been withheld:

1. A technique that divides all samples into k-groups, ideally each having all samples except one.
2. A technique that divides all samples into k-groups, ideally of equal size.
3. A technique that divides all samples into k-groups, ideally each group having the same proportion of target classes.
4. A technique that divides all samples into k-groups, ideally the same target never appearing in both testing and training groups simultaneously.

Pick the option below corresponding to the ordered definitions above:

Label K-Fold, Leave-One-Out, Stratified K-Fold, K-Fold

Leave-One-Out, K-Fold, Label K-Fold, Stratified K-Fold

Leave-One-Out, Stratified K-Fold, K-Fold, Label K-Fold

Leave-One-Out, K-Fold, Stratified K-Fold, Label K-Fold  
correct

Stratified K-Fold, Leave-One-Out, K-Fold, Label K-Fold

**m7q6**

1.0/1.0 point (graded)

There are four teaching assistants in your sister's data science course. Three of them are very sharp and completely understand the material. But one of them, unfortunately, isn't as good with machine learning as the others!

Your sister recorded the following four statements in her notebook, one made by each TA. Can you find the statement made by problematic teaching assistant?

SciKit-Learn provides you with more than one powerful tool for conducting searches over your estimators' parameter values.

Just about all methods called on your GridSearchCV and RandomizedSearchCV classes are actually executed against the highest scoring estimator

When you build a GridSearchCV object, be sure to pass in n\_iter so you can control the number of parameter settings that get sampled.  
correct

When you pass estimator parameters into GridSearchCV, they should be in the form of an array of dictionaries. When you pass them into RandomizedSearchCV, they should be in the form of a single dictionary.

**m7q7**

1.0/1.0 point (graded)

Confusion matrices help you calculate (check all that apply):

How many predictions of a target your model guessed

The number of true positives + false negatives

The number of false negatives + false positives

How many observations per target exist in your dataset

correct