Review1

For `Num` bers:

- Define mean and standard deviation.
- What does a large sd mean?
- What does a small sd mean?
- Show a list of numbers for which the standard deviation is zero

For `Sym` bols:

- There are two terms term1,term2 that are analogous to mean and standard deviation. What are they?
- Given the set of symbols “aaaabbc”, calculate those terms (it is ok to leave the result as fractions).
- Adjust the set of symbols such that term1 changes value.
- Adjust the set of symbols such that term2 increases or decreases.
- Show a list of symbols for which the “term2” is zero.

For each of *TOOL* ∈ (blusterers, classifiers, regressors, and multi-objective optimizers).

- Define what *TOOL* does
- Give a specific example of when you would use *TOOL*.
- (Here are two questions you may not be able to answer… yet)
    - Name a commonly used algorithm for doing *TOOL*;
    - Very briefly, describe how *TOOL* operators.
- Assume data headers can have the special characters;
    - “!” (for “class”)
    - “<” for (“minimize”),
    - “>” for (maximize),
    - “$” (for number)
    - and that anything without any mark is an independent symbol.
    - Assume a data set has five columns with a header `name,age,daysTillDeath, zipCode,income`

        - Write down a header that means we should use *TOOL*.

- Assuming A,B,C,D are the true negative, false negatives, false positives, true positives (respectively) seen
  by a classifier , then

    - define accuracy
    - define recall
    - define false alarm
    - define precision
    - When is “accuracy not accurate”? Give specific values to A,B,C,D where a classifier has a high accuracy yet
      usually misses the target concept (i.e. high accuracy, low recall)
    - When is “precision not precise”? i.e. Give a specific example where a classifier has a very low
      precision, yet still might be considered useful.
    - The following question using the following function.
        - What is the accuracy seen below?
        - What is the recall for “yes”?
        - What is the false alarm rate for “no”?
        - What is the precision for “maybe”?

```
function _abcd(f,i,j) {
  Abcd(i)
                             # want,    got
                             #------    -------
  for(j=1; j<=6; j++) Abcd1(i,"yes",    "yes")
  for(j=1; j<=2; j++) Abcd1(i,"no",     "no")
  for(j=1; j<=5; j++) Abcd1(i,"maybe",  "maybe")
  for(j=1; i<=1; j++) Abcd1(i,"maybe",  "no")
  AbcdReport(i)
}
```

Multi-objective optimization

- define the Boolean domination predicate suitable for two objectives
- Assuming we want to minimize power consumption and cost
    - draw 20 dots on a two-d grid. Mark the Pareto frontier.
        - on that first drawn, draw a tiny square around any dot “*X*” then draw the (much larger)

rectangle indicating which other
dots are dominated by “*X*”

- Make a second drawing (on a new pierce of paper)
  - draw 10 circles from optimizer one and 10 crosses from optimizer two
  - draw the reference frontier

- Define hypervolume, spread, generational distance (GD), inverse generational distance (IGD)
  - On the second drawing, show one distance measurement that would be make for GD but not for IGD
- (Here are two questions you may not be able to answer… yet)
  - define a domination predicate suitable for 3,4,5 objectives.

Review2

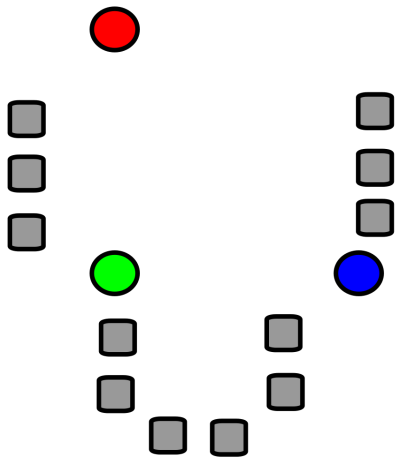
- What are data pre-processors? Can you list three pre-processors in scikit-learn? Use two sentences to explain each of them. Why do we use pre-processors?
- What does the train_test_split function do in scikit learn?
- Can you explain the steps of K-nearest neighbors?
- What are the strategies to split the data with decision tree? (Need to know how to calculate Gini index and information gain)
- What are the differences between Decision Tree and Random Forest?
- Can you explain how Logistic Regression works? What is quantizer? How does multi-class logistic regression work?
- Need to understand and know how to compute confusion matrix, recall, false positive rate, precision, f-measure and g-measure
- How to determine the centroids in KMeans clustering algorithms?
- Explain linear regression. What are simple linear regression and multiple linear regression?
- What are the differences between Decision Tree classification and Decision Tree Regression
- How to evaluate regression models? Can you list two metrics to evaluate regresion models?

Review 3

KNN

- Define
- Here is a distance function that might be used in KNN : $\sum (x-y)^p))^{1/p}$
    - At p=2, does this function have another name?
    - Assuming p=2, what is the distance between (1,0,0) and (0,1,0)? Feel free to leave your answer partially incomplete (just show me how you would do it).
- Explain the following terms: KNN grabs some **smaple** of all the data then applies a **kernel** to summarize then **k** neighbors
- What are some possible values for
    - **kernel**
    - **k**
    - **p**
- If applying KNN to **all** 1000 training examples, what is the value of **sample**
- Explain: “KNN is a lazy learner”
- Explain: “KNN is very slow on large data sets”
- Explain: “KNN can be optimized via clustering”

K-Means



- At each step of k-means, examples label themselves with their nearest centroid. Using the above example, illustrate that labelling step.

- How are the initial K centroids selected (give any one method)
- How are the K centroids updates?
- How does K-means terminate??
- How can K-means optimize KNN? Be specific.

ZeroR

For the following data:

outlook	temperature	humidity	windy	play
rainy	cool	normal	TRUE	no
rainy	mild	high	TRUE	no
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
sunny	mild	high	FALSE	no
overcast	cool	normal	TRUE	yes
overcast	hot	high	FALSE	yes
overcast	hot	normal	FALSE	yes
overcast	mild	high	TRUE	yes
rainy	cool	normal	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	mild	normal	FALSE	yes
sunny	cool	normal	FALSE	yes
sunny	mild	normal	TRUE	yes

- After reading the above data, what does Zero predict for the following? How does it comptue it?

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

- Explain: KNN and K-Means are unsupervised learners but ZeroR is a supervised learner.
- The above data is symbolic. If it were numeric, how would you adjust a Naive Bayes classifier?

Naive Bayes

- For the same data and example as above, what does NB predict? Show all working.
- Explain: “Naive Bayes is **not** a lazy learner”
- Explain why Naive Bayes is “naive”.
- “Naive” Bayes has a low memory footprint and fast incremental update. Why?
- Is Naive Bayes a supervised or unsupervised learner?

Test and split:

- What is the over-fitting problem? Give an example. How can separating data into train and split address that issue?
- What are order effects? Given an example. How can randomizing the order of the data mitigate that problem?
- Define:
    - N-way cross-validation
    - M*N-way cross-validation
- Explain and expand
    - “For a very small data set, do not do 10-way.”
-
    - “For very slow learners (e.g. deep learning), cross-val is hard”
- Explain (definition terms as needed):
    - “Leave one out is an extreme case of N-way cross validation”
    - “Level one out is not recommended for very slow learners”
- Incremental validation
    - Define
    - Explain: “when the learner is slow to update, best to read the data incrementally in large chunks”
    - Explain and fix: “in my incrementally learning, I saw the test case variance increase towards to end of the run”
- Temporal cross-validation:
    - Define
    - Explain: “temporal cross-val avoids the problem where standard n-way uses future data to predict the past”
    - What to do when the data set has no time stamps?

Statistics

Given N learners, an N*M cross-val will generate N*M results per data set. Now we have to check is learner A is better than learnerB.

Ttests:

T-tests check if they can a sample might come from either of two distrubtions.

- Draw normal bell curves for which it is
  - **easy** to show that the sampels are different.
  - **hard** to show that the sampels are different.
- Explain: “t-tests make parametric assumptions about the data”
- Draw two distrubutions that would not be suitable for t-tests

```
DULL=0.147
def cliffsDelta(lst1,lst2, dull = DULL):
     Warning: O(N)^2."""
  n= gt = lt = 0.0
  for x in lst1:
    for y in lst2:
      n += 1
      if x > y:  gt += 1
      if x < y:  lt += 1
  return abs(lt - gt)/n <= dull
```

Cliffs-Delta

- The above code shows the cliff’s Delta effect size test for checking if one list of numbers is different to another
  - Explain this code.
- Explain: “Cliff’s Delta is a non-parametric test”
- What is the computational complexity of the above cliff’s Delta code? How might that complexity be reduced?

Scott-Knott:

- Explain: “Its like a recursive clustering algorithm”
- When splitting treatments, what is the function Scott-Knott is trying to maximize?
- Explain: “When recursing, sometimes Scott-Knott will not descend into a split”
- Explain: “Scott-Knott could use Cliffs-Delta or t-tests or any other statistical test”
- For the following example:

- Sort by medians
- sketch the quartiles (or, as done in class, the 10ths 30ths 50ths 70ths 90ths)
- propose groupings where similar results are placed together

```
RX   10 %   30 %   50 %   70 %   90 %
--   ----   ----   ----   ----   ----
x1 : 0.34,  0.49,  0.51,  0.51,  0.60
x2 : 0.60,  0.70,  0.80,  0.80,  0.90
x3 : 0.15,  0.25,  0.35,  0.35,  0.40
x4 : 0.60,  0.70,  0.80,  0.80,  0.90
x5 : 0.10,  0.20,  0.30,  0.30,  0.40
```

Review 4

Ensembles

Bagging

- What is bagging
- Why can 10 “bags” do better than one?
- What is the connection of bagging to cross-validation?

Boosting

- What is boosting?
- Why can 10 “boosts” for better than one?
- What is the difference between bagging and boosting
- Explain: bagging is inherently parallel while boosting is inherently sequential

Statistics (more)

T-tests

T-tests report the overlap of two normal bell shaped curves. Describe how they might be used in a 10-way cross-val experiments to rank different learners (note: your description should be high-level: no formulas required).

Using the terms standard deviation and mean, describe two such curves that t-tests would find

- easy to distinguish. Draw those two curves.
- hard to distinguish due to some value of the means. Draw those two curves.
- hard to distinguish due to some value of the standard deviations. Draw those two curves.

A t-test is a parametric statistical significance test

- what parametric assumptions re made by the t-test?
- for what kind(s) of curves does the t-test not hold?

Bootstrap

A bootstrap is a non-parametric test of statistically significant difference:

- What is the role of the “test statistic” in the bootstrap?
- What is the role of “sampling with replacement” in the bootstrap?
    - Given the list of numbers [10,20,30,40] write down 5 such samples
- Using the terms “test statistic” and “sample with replacement”, describe the bootstrap (note: your description should be high-level: no formulas required).

Cliff’s Delta

Cliff’s Delta is a non-parametric effect size test:

- What is the difference between an effect size test and a significance test?
- How does Cliff’s Delta sample two lists to see if one is more than trivially different to the other? Use pseudocode
- Explain: non parametric methods are slower than parametric method since they have to using sampling, not summarize , of the data

Clustering:

Distance

- Describe the distance measure presented in lectures and how it can do distances between vectors containing symbolic or continuous values
- Why does that distance metric normalize numbers 0…1 min…max
- How does that metric handle missing numerics?
- How does that metric handle missing symbols?

Mini-batch k-means (MBKM)

- Explain. MBKM works in batches to move centroids less and less.
- Explain: MBKM is a better choice than K-means for very large data sets

KD-trees

- Describe the KD-tree algorithm. Assume ranking via max standard deviation and splits at mean.
- Explain: KD-trees suffer from the curse of dimensionality

Random projections

- Given the distance measure described above
    - describe a random projections clustering algorithm that divides 900 examples into clusters of around size 30
    - Describe how to reduce the curse of dimensionality in KD-trees
- Given a matrix generation algorithm that files columns with gaussians pulled from (mean,sd)=0,1 describe how to reduce
    - a matrix A with m × n rows and columns
    - to a narrower matrix B with m × p rows and columns (p < n)

Aside:

For this kind of RP, it has shown that the Gaussian distribution can be replaced by a much simpler distribution such as

R[i,j] =

- sqrt(3) with probability 1/6,
- zero with probability 2/3,
- and -sqrt(3) with probability 1/6.

Review 5

Clustering

Distance Calculation Tricks

In the following code `i.fun` is some two argument function that returns the distance between two
rows `i,j`. The code illustrates an important trick for reasoning over distance calculations.

- Explain that trick
- Why, when doing recursive random projections might that trick be useful?
- Why, when doing a cross-validation experiment, might that trick be useful?

```
def dist(i,one,two):
    k1,k2 = i.key(one), i.key(two)
    if k1 > k2:
       k1,k2   = k2,k1
    k = (k1,k2)
    if not k in i.cache:
        i.cache[k] = i.fun(one,two)
   return i.cache[k]
```

Pivots for Random Projections

There are at least two ways to do random projections:

1. Gaussian matrix multiplication
2. Random pivots

Explain:

- “Random pivots are preferred when handling symbolic and missing values.”
- “Random pivots can scale to larger data sets than Gaussian matrix multiplication.”

Draw an football-shaped cloud of points on a two-dimensional grid (and that grid have two axes (x,y), each of which run 0…1).

- Using that cloud, explain the Fastmap algorithm for finding two distant points

Draw a second football-shaped cloud of points on a two-dimensional grid that illustrates “the problem of outliers”. Make sure you write
text to explain that problems.

LSR

(This section might require some self study. See here.)

Draw two squares.

- In square one, draw some dots showing a very very strong negative correlation between x and y
- In square one, draw some dots showing a weak positive correlation between x and y
- In square three, draw some dots showing zero correlation between x and y

LSR

- What is “LSR” short for?
- What is the output format of the model found by LSR?
- How could that model be used for classification?

Draw some dots that run from bottom left to top right on a two-dimensional grid (and that grid have two axes (x,y), each of which run 0…1).

- On that plot, show a line that might be added by LSR
- On that plot, show a line that probably would not be added by LSR