CSCI E 82

Exam Review Section - 10th November, 2018

We are giving an exam so that you will have a chance to learn, revisit, and achieve an understanding of the material presented in this course. The exam is open-book, open-computer, open-note but strictly not open-contact with anyone regarding the test questions during or after your test until Monday. Sharing questions or answers during the 48 hour window of the exam is a violation of Harvard policy.

Although the exam allows open materials, the intent is for you to show us what you have learned. It is not an exercise in information retrieval. Consequently, unacceptable answers include:

- Copying text from lecture or section notes
- Taking a screenshot of a lecture slide or section code
- Internet searching and finding some random answer that we never covered
- Internet searching and plagiarizing a discovered answer as your own even with a citation. This is a test of your knowledge and not your internet search prowess

- You will need an internet connection for 2 hours for the real exam that will start 12:01am EST Saturday 17 Nov. and run through 11:59pm EST on Sunday night 18 Nov. -- a 48 hour window.
- Note: You will not have the full two hours if you start after 9:59pm EST on Sunday night.

Problem 1: The following output for a multiple linear regression provides the model. Praise or critique the model based on the findings.

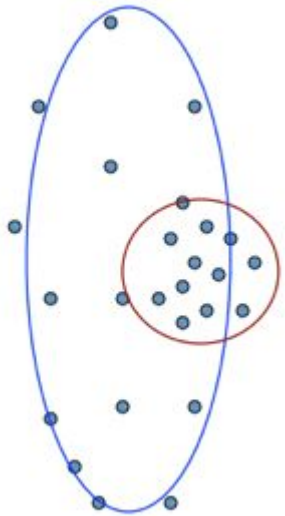
```
Coefficients:
            Estimate  StdError t-value  Pr(>|t|)
(Intercept)  7.000    0.045    <2e-16    ***
x             2.300    0.028    <2e-16    ***
x:y           1.500    2.598    0.433
y             2.940    1.598    0.033     *
x:z          -1.770    0.053    0.007     ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.935e-15 on 98 degrees of freedom
Multiple R-squared:      0.953,    Adjusted R-squared:      0.912
F-statistic: 7.836e+31 on 1 and 98 DF,  p-value: < 2.2e-16
```

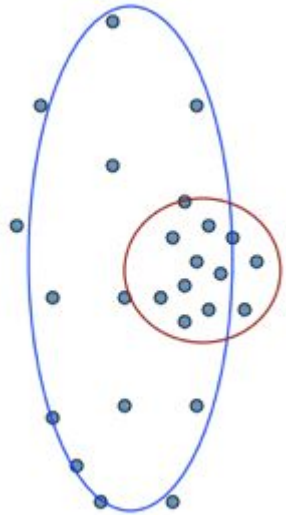
Problem 1:

- Non-significant terms are included
- Hierarchy principle fails with the z not included.
- Adj. R^2 is good.

Problem 2: For a mixture model clustering to work, what parameters need to be estimated for this problem?



Problem 2: For a mixture model clustering to work, what parameters need to be estimated for this problem?



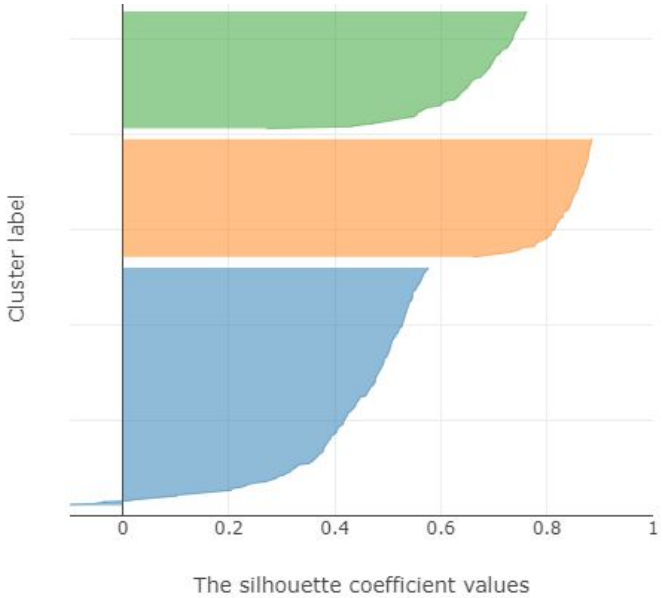
- Parameters for GMM model -
  - Need to model each Gaussian distribution
    - Each 2D Gaussian has a mean(x,y) and a 2x2 covariance matrix
  - Ratio of the two distributions (red vs. blue)

Problem 3: Machine learning often recommends having a training set, validation set, and a test set. Why?

Problem 3: Machine learning often recommends having a training set, validation set, and a test set. Why?

- Training set is for optimizing the parameters.
- Validation set is for assessing models for the training set
- Test set is for assessing the performance

Problem 4: Why is the silhouette method used more frequently for assessing clusters than contingency tables that quantify the predicted vs. actual accuracy?



	Reference Category			
	A	**B**	**C**	**D**
A	22	1	0	0
B	2	2	6	7
C	4	3	17	8
D	0	3	0	18

Problem 4: Why is the silhouette method used more frequently for assessing clusters than contingency tables that quantify the predicted vs. actual accuracy?

The silhouette score is b(i) – a(i) / max{b(i), a(i)} for each point i where it compares the average distance to members of the same cluster and members of the closest other cluster.

a(i): The mean distance between a sample and all other points in the same cluster
b(i): The mean distance between a sample and all other points in the *next nearest cluster.*

Silhouette score is an unsupervised method much like clustering, used when ground truth is not available (intrinsic method) The score is bounded between -1 for incorrect clustering and +1 for highly dense clustering. Scores around zero indicate overlapping clusters. The score is higher when clusters are dense and well separated, which relates to a standard concept of a cluster. The Silhouette score is generally higher for convex clusters than other concepts of clusters, such as density based clusters like those obtained through DBSCAN.

In order for us to use a contingency table, we would have to know the ground truth so it is an extrinsic metric. Other extrinsic methods - homogeneity score, completeness score.

Problem 5: Which method is more efficient for hierarchical clustering: the Reciprocal Nearest Neighbor or Distance Matrix approach?

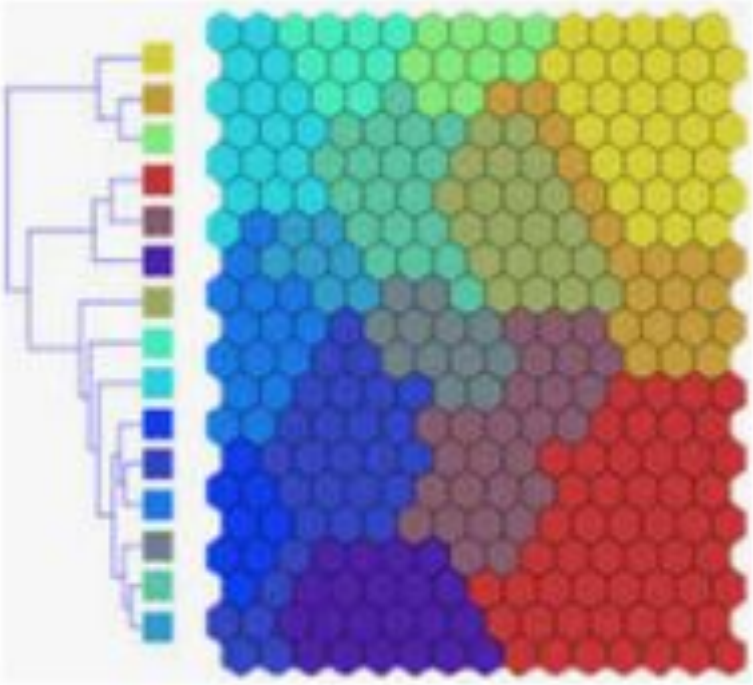
Problem 5) Which method is more efficient for hierarchical clustering: the Reciprocal Nearest Neighbor or Distance Matrix approach?

Distance metric method pre-computes all pairs distances. It iteratively merges them and updates the distances after each merge using the Lance-Williams equations.

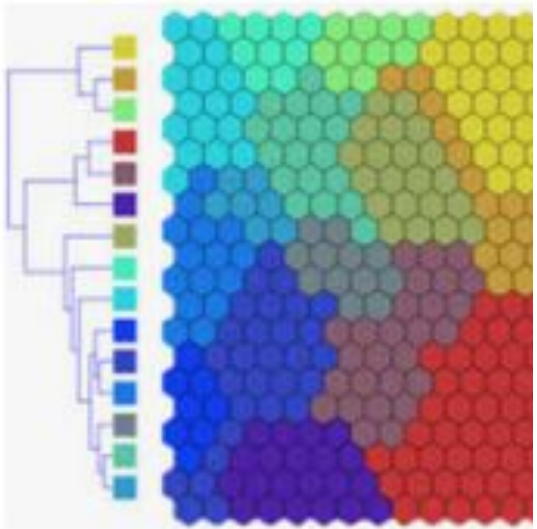
RNN computes distances of each point to others, finds the nearest neighbor, and then repeats that process for that point until it finds a point such that its nearest neighbor is the previous point. It merges these using the Lance-Williams equations, then the process repeats. It doesn't have to store the distances.

Distance metric method is faster but requires a lot of space and updates. It is memory intensive and will only work for moderately sized data sets. The RNN requires many nearest neighbor (1-against-everything) for every iteration which is a lot more computation.

Problem 6) How does the SOM algorithm manage to get similar locations within the SOM map to represent similar entries in the N-dimensional space?



Problem 6) How does the SOM algorithm manage to get similar locations on the SOM map to represent similar entries? [D]



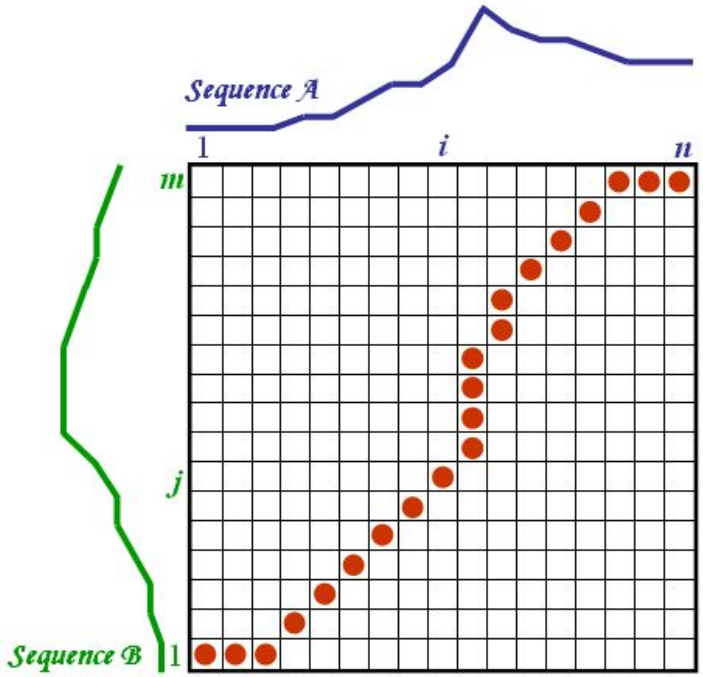
A 2d array of pointers

Every other hexagon cell points to a (cyan) location in N-dimensional space that learns/updates/moves toward a representative location (blue smiles). Not only does it move, but the neighbors in the map (the 2D array) have their representatives also move toward that same location.

The learning algorithm determines the size of the neighborhood that moves, but generally as the algorithm progresses, epoch to epoch, the size of the neighborhood decreases and as does the distance that the representative moves. It is a competitive learning approach like neural nets.

The net effect is that similar regions on the map will be mapped to similar regions in N-dimensional space

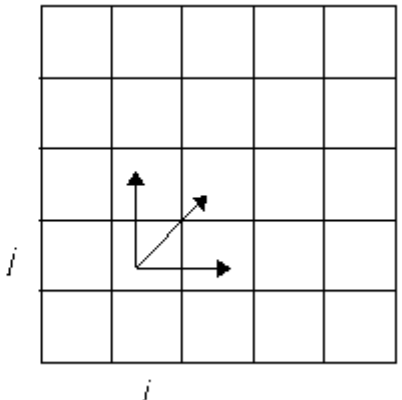
Problem 7) In Dynamic Time Warping (DTW), what do the cells mean and why does the algorithm proceed from one corner to the opposite corner?



Problem 7) In Dynamic Time Warping (DTW), what do the cells mean and why does the algorithm proceed from one corner to the opposite corner? [D]

Cell(i,j) represents the distance between sequenceH(i) and sequenceV(j) so it represents a distance matrix.

The goal is to find the shortest path through the distance matrix. The path starts matching both cells at the start and uses a greedy approach to choose the next closest match going one unit North,NorthEast, or East at each step.



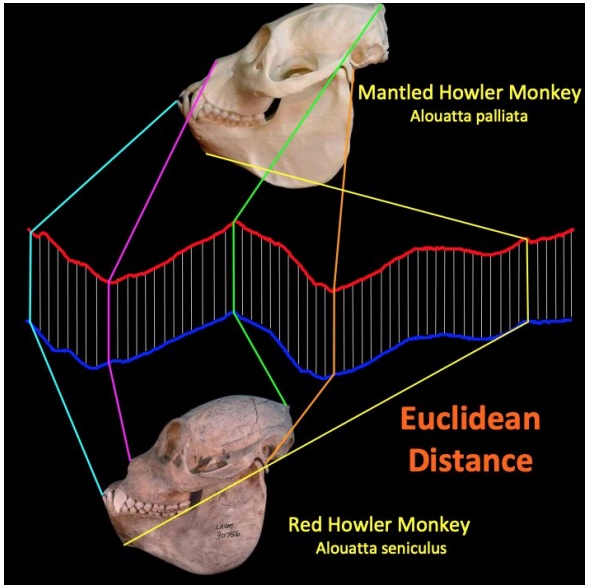
Problem 8) A useful app for hikers and nature lovers is recognizing trees by their leaves. Since color changes significantly with lighting condition, it has been found that the outline or shape of the leaf is a key indication. What would be an effective approach?

Problem 8) A useful app for hikers and nature lovers is recognizing trees by their leaves. Since color changes significantly with lighting condition, it has been found that the outline or shape of the leaf is a key indication. What would be an effective approach? [D]

Represent the outline as a sequence

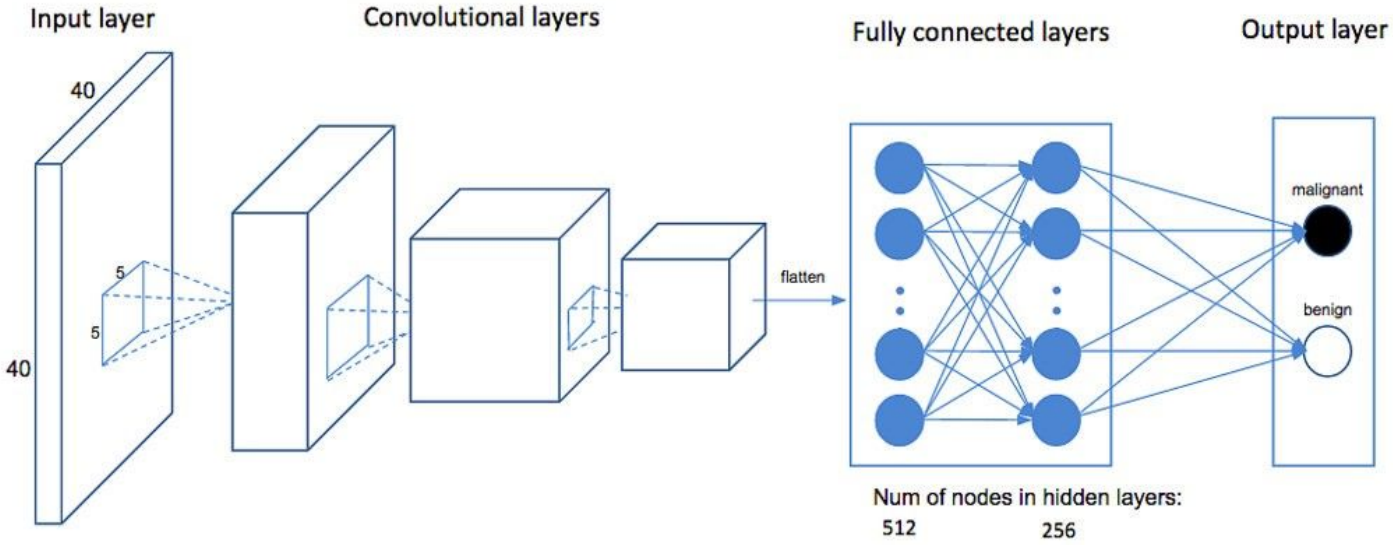
Apply the DTW method

Use K-NN will create a powerful classifier.



Problem 9) Applying deep learning to images is a computationally intensive task. How do CNNs speed up the training process in terms of architecture or approach?

Problem 9) Applying deep learning to images is a computationally intensive task. How do CNNs speed up the training process in terms of architecture or approach?
[D] Through the use of: shared weights, effective optimizers for CNN, batch normalization, max pooling, choice of batch size (convergence/ also memory size and computations). The max pooling step in particular reduces the problem scale substantially at each step shown below on the left..



Problem 10: How does convolution work?

Problem 10: How does convolution work? [D]

- Image convolution applies a small 2D mask to every point of the whole larger image to produce a new image
- In convolution, the mask values are multiplied against the underlying image values. These products are then added to produce the output value. It is essentially a linear operation (review link: here)

Why bother?

- Mimics the local receptive field of visual cortex
- Multiple filters produce a feature map "sandwich" which is one convolutional layer - each map represents a single feature/pattern
- Uses hierarchical feature representations - combine lower level features/patterns (e.g. edges) to make more complex higher level patterns

Problem 11: What is the role of the gradient and learning rate in the gradient descent and how do these interact?

Problem 11: What is the role of the gradient and learning rate in the gradient descent and how do these interact?

The goal is optimization. The gradient/slope determines the direction. The learning rate determines the distance. If you take small steps and re-check the distance then it can be very slow. If you take big steps, then you may miss the optimal or even overshoot. At typical approach is taking larger steps at the beginning with high slopes and then smaller steps as it flattens near the optimal point.

If the region is not convex, then starting at different places may find different local minima. One should try multiple starting points or use a gradient descent optimizer approach such as ADAM used in TensorFlow.
https://stats.stackexchange.com/questions/184448/difference-between-gradientdescentoptimizer-and-adamoptimizer-tensorflow

Problem 12: The training and test data in your AlexNet improve through 30 iterations and then the test data gets worse while the training improves. Explain what is going on and how you would mitigate the problem.

Problem 12: The training and test data in your AlexNet improve through 30 iterations and then the test data gets worse while the training improves. Explain what is going on and how you would mitigate the problem. [D]

What is going on is the network is overfitting after iteration 30.

Solutions—

- Regularize with L1 or L2 penalty on loss function
- Add dropout to one or more convolutional layers
- Add max-norm regularisation to "clip" the weights of neurons in a layer
- Simplify the network - fewer layers, less depth, smaller filters, max-pooling
- Get more data or use data augmentation on the images
- Batch normalization can help, although it is not the main reason for this
- Use early stopping ?

How would this be different if error/accuracy on the training data did not improve?

Problem 13: In regression and most machine learning algorithms, we aim to minimize the variance. PCA searches for the solutions that maximize the variation. Why is this?

Problem 13. In regression and most machine learning algorithms, we aim to minimize the variance. PCA searches for the solutions that maximize the variation. Why is this? [D]

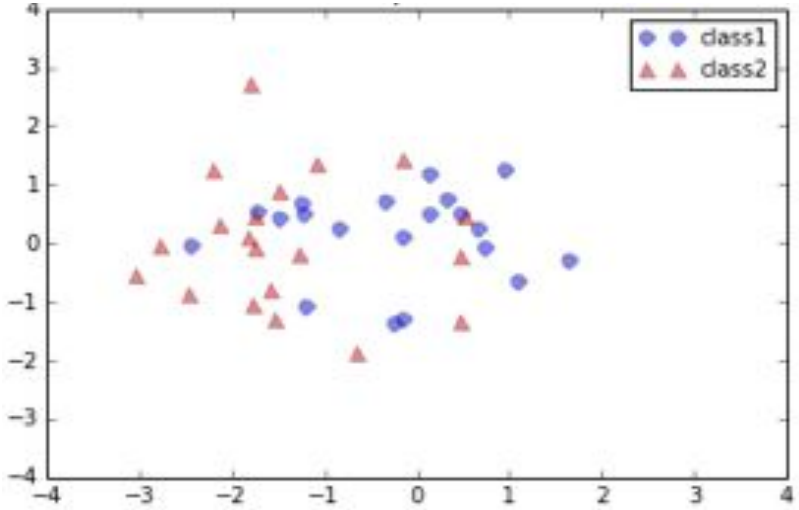
Regression is minimizing the error (residuals).

K-means optimization minimizes the intra-cluster distances.

PCA is selecting axes to capture the maximum range of variation (variance).

(Regression is supervised and PCA is unsupervised. PCA attempts to reduce the dimension with minimal loss of information.)

Problem 14: The following figure shows the PC1 (x-axis) and PC2 (y-axis). What questions should you be asking before making an interpretation?



Problem 14: The following figure shows the PC1 (x-axis) and PC2 (y-axis). What questions should you be asking before making an interpretation? [D]

What fraction of the total variation in the data is modeled by the 2 PCs?

We would like these 2 PCs to represent the largest share of variation in the data.

One should not draw a conclusion from the points in the figure if only a small fraction of the variation in the underlying data is represented.

The fraction of the variation in the data that is explained by the the 2 PCs is the sum of the first 2 |eigenvalues| divided by the sum of all |eigenvalues|.

Problem 15: Name two limitations of using the Euclidean distances to identify outliers?

Problem 15: Name two limitations of using the Euclidean distances to identify outliers?

Only looks at global distances rather than local effects.

Curse of dimensionality problems for high dimensional.

Outlier could related to density rather than distance.

Cannot find an outlier clusters.

Problem 16 How would you construct a bagged KNN algorithm? What benefits would it offer?

Problem 16 How would you construct a bagged KNN algorithm? What benefits would it offer?

Bagging is bootstrap aggregation. Bootstrapping randomly samples the data and generates a model. The aggregation part combines the predictions from multiple models. KNN is the K-nearest neighbor classifier that determines the class based on the nearest neighbors to a given point. One could create an ensemble of multiple samples of the full data set and for each could vote on the class of a given point.

You would have a confidence measure (#votes). It would be less sensitive to the number of features or overfitting in general.

Problem 17 Compare/contrast cross-validation and bootstrapping for assessing classifiers.

Problem 17 Compare/contrast cross-validation and bootstrapping for assessing classifiers. [D]

Cross-validation: Divide the known data into 3-10 distinct pieces or “folds”. For each fold, you train on the rest and test on that fold. You end up testing on the whole known data so every point is assessed once.

Bootstrapping: Sample B (~63.2% of data) bootstrap samples with replacement as training data, and test on the remaining samples. This process is repeated many times, sampling with replacement. While it usually takes a smaller fraction of the known data, you would typically repeat the process *k* times to get unbiased error estimate, so may take longer. Only viable approach when you have a small data set.

Problem 18. You determine that your regression model is overfitting the data. What options are available for this model to mitigate the problem.

Problem 18. You determine that your regression model is overfitting the data. What options are available for this model to mitigate the problem.

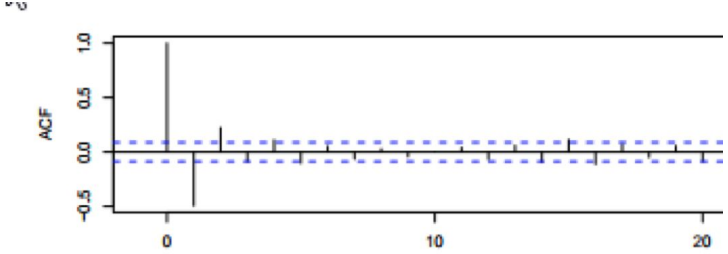
- Use regularization
- Simplify your model by removing terms or interactions
- Get more data

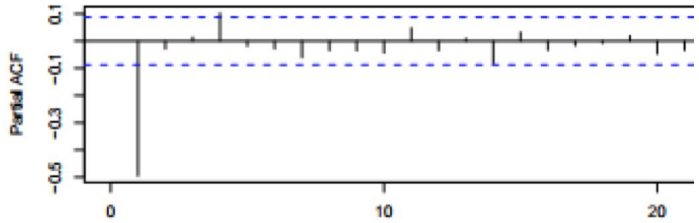
Other things worth trying in general but not necessarily directly related to overfitting:

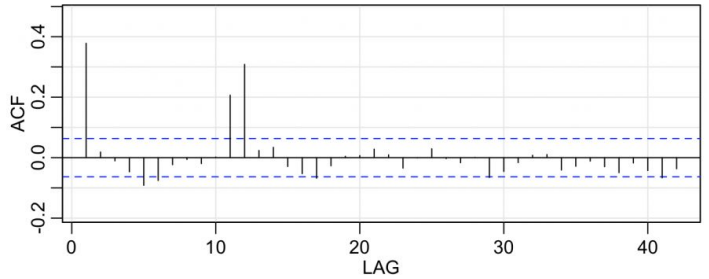
- Check regression diagnostics
- Check for multicollinearity
- Remove outliers (if any)

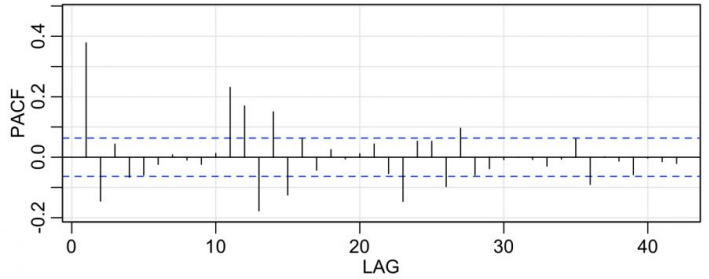
Problem 19) In the ACF plots of time series, what would be the main difference be between an ARIMA(1,0,0) and an ARIMA(1,0,0) x (1,0,0)12?

Problem 19) In the ACF plots of time series, what would be the main difference be between an ARIMA(1,0,0) and an ARIMA(1,0,0) x (1,0,0)12?









The ACF and PACF peaks would not occur at lags of 1 & 2 for ARIMA(1,0,0). The seasonal model would additionally have peaks around 12 & 24.

Problem 20. What are the challenges of using a supervised approach for outlier analysis and how could you get around them?

Problem 20. What are the challenges of using a supervised approach for outlier analysis and how could you get around them?

Usually don't have many outliers so there is a class imbalance issue. One-sided supervised classification is one way around this. (Distance or density based methods have been developed as well but aren't generally supervised.)

SMOTE resamples the data effectively to find a more balanced training set.

Outliers do not necessarily share common features.

Assumes you know what the outliers are.

Problem 21 Word2vec and Latent semantic indexing offer a way of identifying related words but use different approaches. Briefly describe how each achieves this.

Problem 21 Word2vec and Latent semantic indexing offer a way of identifying related words but use different approaches. Briefly describe how each achieves this.

Word2vec is used to produce word embeddings which represents words in dense numeric vector format (unlike TFIDF where were word vectors are quite sparse). It is built using the surrounding several words so it is leveraging the local context to identify similarities.

LSA/LSI (TruncatedSVD) helps transform the sparse document x word matrix into a lower dimensional space, which can be further used to visualize or cluster the dataset. It is using the co-occurrence of words between documents rather than the local context. Similar words will be mapped into the same projection space.

Problem 22. Isolation trees generally continue until there is one node per leaf. Decision trees are usually built and pruned back so that leaves are no longer not a single class. Random forests continue until a set depth but generally are never pruned. Explain the differences.

Problem 22. Isolation trees generally continue until there is one node per leaf. Decision trees are usually built and pruned back so that leaves are no longer not a single class. Random forests continue until a set depth but generally are never pruned. Explain the differences.

- Isolation trees are for outliers. The algorithm builds an ensemble of trees using random cuts for randomly selected features. The lower the average depth for a data point (closer to the root) —the more outlier-like. Since random cuts are made, many trees are needed. (If you prune isolation tree, you won't be able to get the actual depth of the tree - it relies on determining outlier based on depth of the tree.)
- Decision trees are prone to overtraining and are pruned. This is the bias variance trade-off.
- Random forests using bagging and so do not suffer overtraining. They therefore do not need to be pruned. The randomness is based on the bootstrap selection of data points and random subset of features at each node.

Problem 23 ）What is variance inflation and what should be done if it occurs?

Problem 23) What is variance inflation and what should be done if it occurs? [D]

Caused by correlated (identical, or linear scaling) of features in regression. E.g. what happens if you include two copies of the same feature in a regression.

In regression, we estimate each coefficient and its standard error (aka variance) - these become nonsense in this case. Because of the features are equivalent or similar, the weights can take a range of values as they balance out. This range of values causes a large variance of the weights.

What to do about it?

- Remove the superfluous highly-correlated features - keep only the one most correlated with y
- Apply PCA to the features and use the resulting PCs as new features. PCA produces uncorrelated (orthogonal) features that can be used in regression.

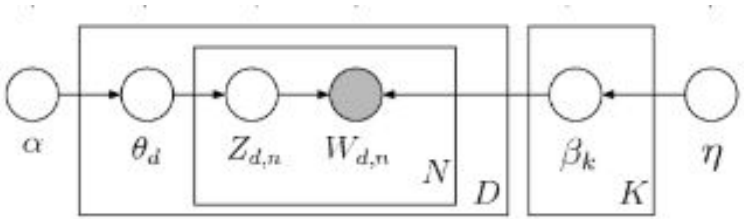
Problem 24) What do the eta (η) and alpha (α) parameters control in the LDA model?

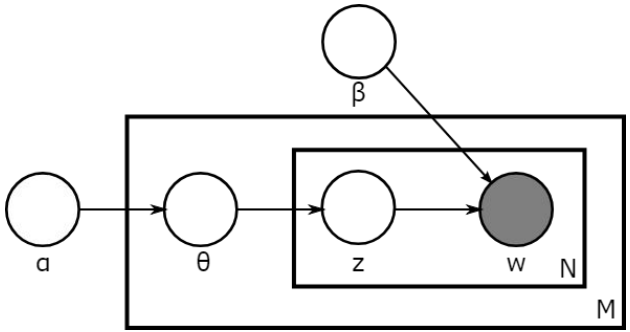
Problem 24) What do the eta (η) [often also beta] and alpha (α) parameters control in the LDA model? [D]

Beta or eta (η) - Choice influences the number of topics within a document. If η is high, then each document contains all topics - and the documents are more similar

alpha (α) - Choice influences the number of words within in a topic. If α is high, then each topic contains a mixture of most words - and the topics are more similar

More complicated details: these are two hyper-parameters in the prior Dirichlet distribution which characterizes how words and topics are assigned.





Problem 25) Both linear regression and Latent Dirichlet Allocation were introduced using a generative model. What was the purpose of using the generative model?

Problem 25) Both linear regression and Latent Dirichlet Allocation were introduced using a generative model. What was the purpose of using the generative model?

- Linear regression: Defined linear model as a line containing all points and vertically shifted points off the line using a random value.
- LDA: Defined documents as a random collection of topics. Defined topics as a random collection of words.

The algorithms for these two generative models are based on these assumptions.

Process is Model → data :
start with a model (line & #points or mixture topics & # words)
→ define parameters
→ randomly create data

Problem 26) Naïve Bayes makes an assumption of conditional independence. Comment on this assumption for use in classifying text.

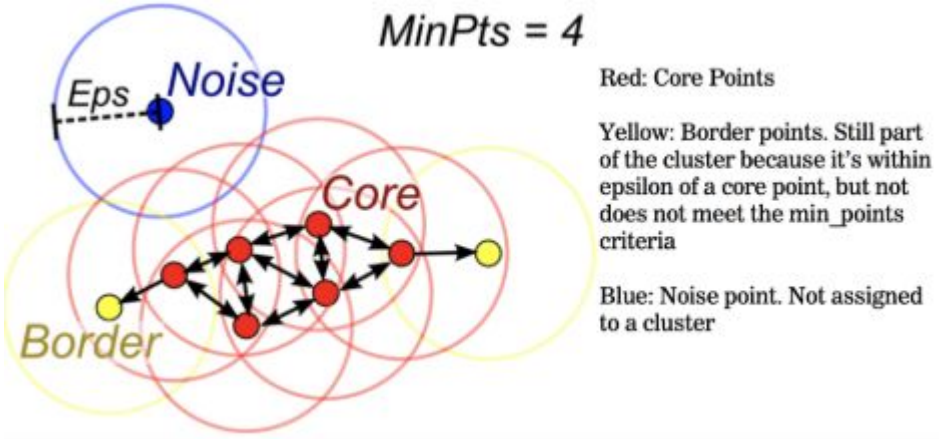
Problem 26) Naïve Bayes makes an assumption of conditional independence. Comment on this assumption for use in classifying text.

Lose all the context. Instead of looking for “hello world” it treats each as a separate BOW essentially with the frequencies of “hello” and “world” separately and no context.

Problem 27) How does DBSCAN define dense regions and how does it use them to form clusters?

Problem 27) How does DBSCAN define dense regions and how does it use them to form clusters?

1) Defines epsilon radius around points and counts points within
2) Core points are centers of these epsilon radius if contain minPts
3) Points are reachable if one is core and other within epsilon radius
4) Connected regions are joined if there is a path of ‘reachable’ points between them



Problem 28) What is TF-IDF and what potential purpose does it serve in clustering documents?

Problem 28) What is TF-IDF and what potential purpose does it serve in clustering documents?

Simple Term-Frequency just gives weight to frequently occurring words in documents. But not all words are created equal.

Term Frequency – Inverse Document Frequency is a weighted version such that rare words are weighted more than commonly occurring words. It provides a different weighting to the words that is useful.

For example: In a search, the weighting of rare words in a query is much more important in finding the correct document.

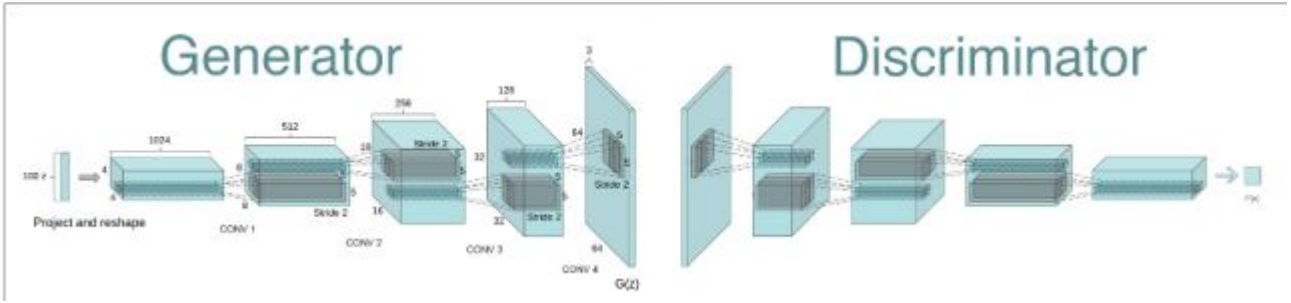
Problem 29) What are the roles of the inspector and the generator in a GAN?

Problem 29) What are the roles of the inspector and the generator in a GAN? [D]

The generator is the network that creates the labelled "fake" data by randomly mapping from the latent space of images to an artificial image (decoder step)

The inspector (or discriminator) is the network that, confronted by the real data and the generated "fake" data, must make a prediction of whether each is a real image from the training dataset or an artificial image created by the generator.

The two networks can be considered adversaries, and the GAN trains to make each of them a better adversary (better at both generating and detecting "fakes")



Problem 30. APRIORI algorithm with the following itemset. Assume 2 is the frequent support.

Item sets: X,Y X,Y,Z, Y,Z, W, X, Y, Z, X,Z W,Y,Z

1-itemsets: W=2 X=4 Y=5 Z=5

2-itemsets: W,X = 1 W,Y = 2 W,Z = 2 X,Y = 3 X,Z = 3 Y,Z = 4

3-itemsets: W,Y,Z = 2 X,Y,Z = 2

Why were only these two 3-itemsets considered?

Why were only these two 3-itemsets considered? [D]

To create a j-itemset, one joins the j-1 itemsets together such that:

- Only those above the minimum frequent support are used
  - So the 2-itemset W,X is ignored.
- The first j-2 items must match and differ only in the j-1 index.
  - The W's match and differ with Y and Z so these merge
  - X's match and so Y and Z differ
  - Y,Z is alone so nothing to merge with

- Apriori property: Subsets of frequent itemset must also be frequent itemsets
- What if frequent support was 3?