List of Test Questions for the first part of the test.

1. What kind of patterns do you search for with data visualization
   A> Clusters , Outliers, other non-linear or linear patterns.
2. Let X be a $n \times p$ data matrix with n observations and p variables. Show how to use the SVD to calculate the principal components of the data.
   VDDV^T $S = V D^2 V^T$
3. What do you need to assume about n and p in prob. 4? Suppose that n < p, can you figure out how to calculate the principal components of X.
   $n \geq p$ $S = V D V^T$ use pcs to compress the feature less than n,
4. Supposed we have a binary response Y and a vector of predictors X. How would you define a linear classifier? $h(x) = x^T \beta + \beta_0$ $\begin{cases} y=1 & \text{if } h(x)>0 \\ y=-1 & \text{otherwise} \end{cases}$ $Y = X\beta + \varepsilon$ $\hat{\beta} = (X^T X)^{-1}(X^T Y)$
5. Is it possible that there might be a *biased* estimator with lower MSE than the Least squares estimator? Yes Gaussian-Markov implies LSE has lowest MSE of all linear estimators with no bias but biased estimators could e.g. trade a little bias for larger reduction in variance.
6. Define the method of "leave one out" cross-validation
   for every point $i = 1 \ldots n$, train the model on every point i, compute the test error on the held out point, average the test errors
7. Define the method of "ten fold" cross-validation
   A. Split data into 10 groups evaluate the response at each group while omitting the group to fit the prediction model.
8. Define what we mean by partitioning the data into training and testing sets.
   A. Training set is used to fit the prediction model and the testing set is used to evaluate the performance of the model
9. Point out the difference between "training and testing sets" and cross-validation methodology in order to evaluate the performance of a method.
   A. In CV all the data is reused in predicting other data while in training – testing no data is reused.
10. Give the constraint to the least squares estimating equations which produces the ridge regression estimator. $\sum_{j=1}^{p} \beta_j^2 \leq s$ $\sum_{j=1}^{p} \beta_j^2 \leq s$
11. Give the constraint to the least squares estimating equations which produces the LASSO estimator. $\sum_{j=1}^{p} |\beta_j| \leq s$ $\sum_{j=1}^{p} |\beta_j| \leq s$
12. Give the constraint to the least squares estimating equations which produces the elastic net estimator. $\lambda \sum_{j=1}^{p} (\alpha \beta_j^2 + (1-\alpha)|\beta_j|) \leq s$ $\lambda \sum_{j=1}^{p} (\alpha \beta_j^2 + (1-\alpha)|\beta_j|) \leq s$

13. Describe the method of Principal components regression.

1. Calculate principal component variables 2. Use principal comp as predictors for the regression model

14. Describe the method of Partial Least Squares regression giving the first iteration.

Do individual regression for each predictor and calculate the linear combination of the slopes. This gives the 1st PLS component

15. Does Least Squares and/or PLS regression require that you to have less variables than observations? Explain.

LS yes but not PLS PLS follows steps, LS depends on the matrix

16. Define cubic spline.

Is a piecewise third degree polynomial that passes by a set of points with continuous 1st and 2nd derivatives.

17. What are the basis functions for a cubic spline?

$1, X, X_2, X_3, (X-b_i)_{3+}$ for i=1…k $1, X, X^2, X^3, (X-b_i)_+^3, i=1,\dots k$

18. What is the penalty criteria for the method of smoothing splines?

$\lambda \int (f''(t))^2 dt$

19. What are the basis functions for natural splines?

$$N_1(X) = 1$$

$$N_2(X) = X$$

$$N_{k+2}(X) = d_k(X) - d_{K-1}(X), \quad \text{where}$$

$$d_k(X) = \frac{(X-\xi_k)_+^3 - (X-\xi_K)_+^3}{\xi_K - \xi_k}$$

20. Give the form of the linear discriminant analysis method for predicting a response Y with three possible values {"1","2","3"}.

1. all Σ's of three groups are equal
2. Calculate distance from y to the center of each group $L_i(y) = a_i' y + a_{i0}$
3. assign y as the tag of the closest group assign y to largest $L_i(y)$

21. In preprocessing the data for clustering, give some available transformations to improve the scale of the data.

1 make sdev equal. 2. Normalize the inverse sqrt of the common cov matrix 3. Normalize each cluster by the common within cluster inverse sqrt covariance

22. What algorithms did you study for cluster analysis and what are their differences? Hierarchical clustering, k-means, PAM and model-based clustering. Only the first one assumes k unknown. Model based assumes a mixture of normal distributions, so there exist k clusters in the population. k-means and pam minimize WCSS, non-robust and robust respectively.

23. Give some examples of hierarchical clustering algorithms and describe their distances. Single linkage (closest) average linkage(average distance) complete linkage (longest distance) Ward is ~~R_2~~ R^2 distance,

24. Define Euclidean distance d(x,y) for x,y p-dimensional vectors.
Sum($(x_i - y_i)_2)_{1/2}$ $\sum_{i=1}^{p} \sqrt{(x_i - y_i)^2}$

25. Define Manhattan distance d(x,y) for x,y p-dimensional vectors.
Sum($|x_i - y_i|$) $\sum_{i=1}^{p} |x_i - y_i|$

26. What is the Ward method for hierarchical clustering?
Uses the most improvement on ~~R_2~~ R^2 to define the hierarchical tree

27. What is singe linkage?
Joints the closest clusters by the closest pair of observations

28. What is the difference between unsupervised classification and supervised classification?
Unsupervised you don't observe the response variable. Supervised you observe the response.

29. Describe briefly the methodology of k-means clustering.
Find the configuration which Minimizes the within cluster sum of squares for a given number of clusters k. Start with an initial configuration and move to better configurations by exchanging points between clusters.

30. What are the criteria that we minimize with k-means clustering?
the within cluster sum of squares

31. Describe the algorithm for k-means clustering and discuss the convergence.
Start with an initial configuration and move to better configurations by exchanging points between clusters. Stop when converged to a local minima. Repeat with new initial configuration as many times as needed to improve the convergence to a local minimum

32. Describe briefly the methodology of PAM clustering.
Minimize the absolute deviations with respect to the cluster medians.

33. Give the silhouette value for an observation.
$S(X_i) = (d2(X_i) - d1(X_i)) / \text{Max}(d2(X_i), d1(X_i))$
$d2(X_i)$ = Ave distance to closest cluster. $d1(X_i)$= Ave distance to own cluster

$$S(X_i) = \frac{d_2(X_i) - d_1(X_i)}{\text{Max}[d_2(X_i), d_1(X_i)]}$$

34. Give the silhouette statistic for evaluating a clustering configuration.
Ave ($S(X_i)$) where $S(X_i)$ is the silhouette value for X_i .

35. What is the basic idea behind the recursive partitioning algorithms?
Split the nodes into buckets until we span the entire data into homogenous buckets.

36. Give the main steps of the Cart algorithm.
1. Build the tree until nodes are too small according to stopping rule.
2. Prune the tree as much as needed using CV or Mallow's CP statistic.

37. Give the CART splitting criteria for continuous response with equal variances.
(Nl sigma^2L + Nr Sigma^2R) / (Nl + Nr)

$$\frac{N_L \sigma^{2L} + N_R \sigma^{2R}}{N_L + N_R}$$

38. Give the entropy index splitting criteria for CART with binary response.
P_L ($-p_{0L} \log(p_{0L}) - p_{1L} \log(p_{1L})$) + P_R .....

39. Give the gini index splitting criteria for CART with binary response.
p_L (p_{0L} p_{1L}) + p_R (p_{0R} p_{1R})

40. Explain the pruning step of recursive partition.
Take terminal split and compare the CV residual with and without split. If better without splitting, then delete split. Repeat until no more pruning is possible.

41. What is a random forest?

Algorithm which generates a forest of trees and takes the average predicted values as the prediction. Each tree is calculated on a bagged sample. The in-bag sample is used to fit the tree and the out -of-bag sample is used to estimate the prediction.

Each split is done with a random subset of predictors.

42. What is the standard number of variables that are sampled at a node of the random forest algorithm?

If you have p variables => sqrt(p)

43. What is the in-bag sample of random forest?

Bootstrap sample with deleted repeats.

44. What is the out-of-bag sample of random forest?

Remaining observations outside the bootstrap sample

45. Give the Steps of the random forest algorithm.

Bag in-bag sample. build the tree. Estimate prediction on out-of-bag sample. Final prediction is average of all out-of –bag predictions.

46. What does it mean bagging?

Sampling n observations out n with replacement and deleting repeats

47. Give the basic steps of the Boosting algorithm.

Built a sequence of trees. At each step increase the weights of the observations misclassified by the previous tree and decrease the weights of the observations classified correctly. Repeat until misclassification rate above 50%. Choose the average of the 1 thru k trees that has the least misclassification error.

48. Define SVM and give the SVM classifier equation.

SVM maximizes the margin seperation between two classes

$f(x) = \sum_i a_i y_i (x_i^T x) + b$

49. Define the margin of a Support Vector Machine classifier and give its value as a function of beta the vector parameter of the SVM classifier equation.

A> Separation between the two classes. 2/||beta||

$\frac{2}{\|\beta\|}$

50. Define the Kernel method for function approximation

A>

$$f(x) = \sum_m \alpha_m K(x, y_m)$$

51. What is the basic structure of a neural net Input layer, hidden layer, Output layer
52. Give the sigmoidal function and method. Sig(t)=1/(1+exp(-t)) $Sig(t) = \frac{1}{1+e^{-t}}$
53. Explain the Softmax method :

$$g_k(T) = \frac{e^{T_k}}{\sum_{l=1}^{K} e^{T_l}} (SOFTMAX)$$

Minimize

$$R(\theta) = \sum_{k=1,i=1}^{K,N} (y_{ik} - f_k(x_i))^2$$

$$R(\theta) = -\sum_{k=1,i=1}^{K,N} y_{ik} \log f_k(x_i)$$

I am interested in developing resampling and bootstrapping methodology that can be applied for the analysis of data that is generated by complex processes. In order to perform standard methodology analysis / data mining, many of these problems require multiple assumptions that are unrealistic and not easy to work out.

Resampling methods have been developed with the goal of solving many of these issues but the methods must be tailored to the specific applications, and this often requires new research and new methodology.

One example of this is the methodology of target estimation for bootstrapping. This is a resampling method that I developed for applications to imaging in computer vision, biology and other areas.