Review Questions

Signal Data Science

Linear regression

- Is linear regression a supervised or unsupervised method?
- What does "least squares" refer to in the phrase "ordinary least squares regression"?
- What does it mean for the coefficient of a feature to be equal to k in a linear regression model?
- How can you use categorical variables as predictors in a linear regression model?
- If you have a categorical variable which can take on one of k values, how many indicator variables should be created from that categorical variable?
- What does R^2 represent?
- What is the difference between R^2 and adjusted R^2?

Assumptions

- What is one assumption made by ordinary least squares regression?
- What assumption is made by ordinary least squares regression about the predictor variables?
- What assumption is made by ordinary least squares regression about the mean of the error terms?
- What assumption is made by ordinary least squares regression about the variance of the error terms?
- What assumption is made by ordinary least squares regression about the correlations between the error terms?
- What assumption is made by ordinary least squares regression about the correlations between the error terms and the predictor variables?
- What assumption is sometimes made by ordinary least squares regression about the distribution of the error terms?

Residuals

- What is a residual?

- After fitting a linear regression model, how can its residuals be calculated?
- How would you plot the residuals of a single-variable linear regression?
- What is one example of information you can get from plotting the residuals of a linear regression?

Cost functions

- What is a cost function?
- What is one example of a cost function?
- What is the minimum value of these cost functions?
- What is the maximum value of these cost functions?
- What cost function is used for ordinary least squares regression?
- Will you get equivalent results using the RMSE and SSE cost functions?
- Will you get equivalent results using the SSE and MAE cost functions?
- Why might you use the MAE instead of the RMSE in a regression model?
- How could you modify the cost function to combat overfitting?
- What is an example of a situation when you might want to give different points in your training data different weights (representing their importance to the model)?
- How would you assign points in your training data non-uniform weights in an ordinary least squares regression?

Sources of concern

- What is the difference between collinearity and multicollinearity?
- What is one example of a real-life situation where you might have perfect multicollinearity?
- What is one example of a real-life situation where you might have imperfect multicollinearity?
- If there is significant multicollinearity between the predictor variables, what do you expect to happen to the coefficient estimates of a linear model as you add and remove predictor variables?
- How might significant multicollinearity make it more difficult to interpret the results of a linear regression?
- If there is significant multicollinearity between the predictor variables, what might the statistical significance of each predictor variable look like?
- You run a linear regression and the response variable is perfectly predicted. What is one possible explanation?
- You run a linear regression and some of the coefficients are absurdly high. What is one possible explanation aside from multicollinearity?

Regularization

- What is the L^1 norm of a vector?
- What is the L^2 norm of a vector?
- What is L^1 regularization?
- What is L^2 regularization?
- Will an unregularized or regularized linear model have a lower error on training data?
- What is one advantage of using a regularized linear model?
- What is one disadvantage of using a regularized linear model?
- What is the major difference in the outcomes obtained by using L^1 or L^2 regularization?
- Why do we use L^1 regularization instead of $L^{1/2}$ or L^0 regularization?
- How can L^1 and L^2 regularization be combined?
- What is one interpretation of regularization in terms of Bayesian modeling?

Resampling

Overfitting

- What is one disadvantage of having more columns than rows in your dataset?
- What is one concrete example of overfitting?
- What is one way to deal with overfitting?

Cross-validation

- What is n-fold cross-validation?
- What is the purpose of n-fold cross-validation?
- What values can n take on in n-fold cross-validation?
- What is the special name used for the case when n is maximal?
- What is an advantage of using a higher value of n in n-fold cross-validation?
- What is a disadvantage of using a higher value of n in n-fold cross-validation?

Bootstrapping

- What is a bootstrapped sample?
- What is the size of a bootstrapped sample?

- Approximately what proportion of the original dataset shows up at least once in a bootstrapped sample?
- Suppose that we generate many bootstrapped samples of a dataset. We can estimate the generalization error by training a model on the original dataset and making predictions on the bootstrapped samples, or by training a model on each bootstrapped sample and making predictions on the original dataset. Which method is preferred?
- If we train a model on each bootstrapped sample and make predictions on the original dataset, approximately two-thirds of the "test" data will already have been seen by the model. How can we account for this?
- How can bootstrapping be used to estimate parameters of probability distributions?

Logistic regression

- What does the "binomial" in "binomial logistic regression" mean?
- What assumption is made about the distribution of the response variable in a logistic regression model?
- What is being linearly modeled in logistic regression?
- What is the minimum value of a log odds?
- What is the maximum value of a log odds?
- What assumption is made about the log odds associated with the probabilities of the Bernoulli distributions in a logistic regression model?

Comparison to discriminant analysis

- What generative assumption is made by linear discriminant analysis?
- Which classification method makes stronger assumptions: logistic regression or linear discriminant analysis?

Maximum likelihood estimation

- What is a likelihood function?
- What is the minimum value of a likelihood function?
- What is the maximum value of a likelihood function?
- What does the area under a probability density function need to be equal to?
- What is the function for the probability density of a uniform distribution from a to b?
- What is the likelihood function for a uniform distribution from a to b and a single observation y?

- A random variable is 1 with probability p and 0 otherwise. What is the likelihood function for a single observation y?
- A set of independent random variables are 1 with probability p_i and 0 otherwise. What is the likelihood function for a set of observations y_i?
- What numerical method is typically used to calculate the probabilities which maximize the likelihood function for logistic regression?
- How can the maximum likelihood estimation of a logistic regression model be interpreted in terms of minimizing a cost function?

Visualization

- What is sensitivity?
- What is specificity?
- What variable is plotted on the y-axis of an ROC curve?
- What variable is plotted on the x-axis of an ROC curve?
- What ROC curve would correspond to a completely random classifier?
- What is the AUC?
- What is one usage of the AUC?
- What is one way to visualize and interpret log odds ratios corresponding to predictions made by a logistic regression model?

Multinomial logistic regression

- What does the "multinomial" in "multinomial logistic regression" mean?
- Briefly describe how multinomial logistic regression differs from standard binomial logistic regression.

Principal component analysis

- Is principal component analysis supervised or unsupervised learning?
- What is the first principal component of a dataset?
- What is the second principal component of a dataset?
- How many principal components does a dataset have?
- What are the "loadings" of a principal component?
- What are the "scores" of a principal component?
- What is a linear combination?
- How can you calculate the scores of a principal component from its loadings?
- What is a geometric interpretation of principal component analysis?
- What technique from linear algebra is typically used to calculate the principal components of a dataset?

- Principal component analysis produces an "eigenvalue" associated with each principal component. What is one interpretation of these eigenvalues?

Visualization

- What is one way to visualize the results of principal component analysis?
- What information can principal component analysis provide you?
- What is one way to visualize the eigenvalues of principal component analysis?

Principal component regression

- What is principal component regression?
- What is one advantage of principal component regression?
- What is one disadvantage of principal component regression?
- How can you quickly determine a good number of principal components to "keep" for principal component regression?
- What is one reason why principal component regression might have much less predictive power than ordinary least squares regression even if the first few principal components explain most of the variation in the data?

Clustering

- Is clustering supervised or unsupervised learning?
- What is one way to measure how "real" a cluster is? *I.e.*, whether or not it's a random fluke.
- What is one reason why clustering is challenging with high-dimensional data?
- What is one way to overcome the curse of dimensionality?

K-means clustering

- Is K-means clustering a generative or discriminative model?
- What is the minimum value of K?
- What is the maximum value of K?
- What are the steps of the K-means algorithm?
- What is one advantage of using K-means clustering?
- What is one disadvantage of using K-means clustering?
- Where does the instability of the results of K-means clustering come from?

- What is one way to account for the instability of the results of K-means clustering?

Minimizing variance

- What quantity is minimized by the K-means algorithm?
- How is the distance between points calculated in the K-means algorithm?
- Can you use other distance metrics with the K-means algorithm?
- As K increases, what happens to the within-cluster variance?
- What is one way to choose a value of K?
- What is the general algorithm used to for K-means clustering called?
- What is the "expectation" step of the K-means algorithm?
- What is the "maximization" step of the K-means algorithm?

Gaussian mixture models

- What is a Gaussian function?
- What probability distribution is represented by a Gaussian function?
- Why is the Normal distribution so commonly used?
- Is a Gaussian mixture model a generative or discriminative model?
- What generative assumption is made by Gaussian mixture models?
- What is the general algorithm used to fit a Gaussian mixture model called?
- Why do people sometimes call K-means clustering "hard" and Gaussian mixture models "soft"?

Decision trees

- What does the acronym CART stand for?
- Intuitively, how does a decision tree work?
- What is a "leaf" of a decision tree?
- In a regression tree, how is a prediction made at a leaf of the tree?
- In a classification tree, how is a prediction made at a leaf of the tree?
- What is one advantage of using a decision tree?
- What is one disadvantage of using a decision tree?
- What is one way to combat overfitting in a decision tree?

Bagging

- What does "bagging" stand for?
- How can bootstrapped samples of a dataset be used to improve decision tree performance?

- How can you estimate the generalization error using models trained on bootstrapped samples?

Random forests

- How do random forests relate to bagging?
- What part of a random forest is random?
- What purpose does the randomness of a random forest serve?
- What is the single hyperparameter of a random forest?
- What is the minimum value of the single hyperparameter of a random forest?
- What is the maximum value of the single hyperparameter of a random forest?
- What is typically considered to be a good default value for the single hyperparameter of a random forest?
- What is one advantage of using a random forest?
- What is one disadvantage of using a random forest?
- How do the trees of a random forest depend upon each other?
- How can the training of a random forest be parallelized across multiple processors?
- How is the standard measure of variable importance for a random forest calculated?

Gradient boosted trees

- What is an intuitive explanation of gradient boosting?
- What are the three hyperparameters of a gradient boosted tree?
- What is one advantage of using a gradient boosted tree?
- What is one disadvantage of using a gradient boosted tree?
- How do the trees of a gradient boosted tree model depend upon each other?
- How can the training of a gradient boosted tree be parallelized across multiple processors?
- What is the name of the most popular software package for parallelized training of gradient boosted trees?

Recommender systems

- What is one example of a problem for which you would use a recommender system?
- What is one example of explicit data collection?
- What is one example of implicit data collection?

- What is one example of a situation where you might want to recommend an already-seen item to a user?
- What is one example of a situation where you might not want to recommend already-seen items to a user?
- Should recommender systems take into account a user's entire history?
- Aside from optimizing solely for recommendation accuracy, what are some other factors a recommender system should take into account?

Collaborative filtering

- What is one concrete, real-world example of collaborative filtering?
- What is the fundamental assumption made in collaborative filtering?
- What is one advantage of using collaborative filtering?
- What is the scalability problem?
- What is one way to overcome the scalability problem?
- What is one disadvantage of collaborative filtering which has not already been mentioned?
- What is the name of one algorithm for performing collaborative filtering?

Content-based filtering

- How does content-based filtering differ from collaborative filtering?
- What is one concrete, real-world example of content-based filtering?
- What is one advantage of using content-based filtering?
- What is one disadvantage of using content-based filtering?

Hybrid methods

- What is one way to combine the results of multiple recommender systems?
- What is the cold start problem?
- What is one way to overcome the cold start problem for users?
- What is one way to overcome the cold start problem for items?
- If the vast majority of a website's users don't have accounts and you can't use cookies to track individual users, how would you design a recommender system?

Advanced topics

k-Nearest Neighbors

- What is *k*-Nearest Neighbors?

- What is one advantage of k-Nearest Neighbors?
- What is one disadvantage of k-Nearest Neighbors?
- When would k-Nearest Neighbors work well?
- How would you extend k-Nearest Neighbors so that closer points affect the model's predictions more than points farther away?

Support vector machines

- What does it mean for two classes to be linearly separable?
- In a linearly separable binary classification problem, what is the margin?
- What is the maximum-margin hyperplane and why is it important?
- How do we fit a linear SVM if the two classes of data are not linearly separable?
- What is a support vector?
- What is a dot product?
- What is a kernel function?
- What is one example of a kernel function?
- What is another name for a radial kernel function?
- What is one example of a situation where you would want to use a linear kernel?
- What is one example of a situation where you would want to use a radial kernel?
- What is one example of a situation where you would want to use a polynomial kernel?

Natural language processing

- What is an n-gram?
- How are n-grams used?
- What is the fundamental assumption of a naive Bayes classifier?
- What is the classic example of a naive Bayes classifier?
- What is sentiment analysis?
- What is topic modeling?
- What is the generative model proposed by latent Dirichlet allocation?
- Given a large dataset of books, how would you organize them into different genres?
- What is one way to represent a document as a vector?
- What is one appropriate distance metric to use for comparing document similarity?
- Is there any special reason to use cosine similarity instead of Euclidean distance for comparing document similarity?

Parametric vs. nonparametric models

- What is a parametric model?
- What is one example of a parametric model?
- What is a nonparametric model?
- What is one example of a nonparametric model?
- How do parametric and nonparametric models differ in their assumptions?
- Is a linear SVM a parametric or nonparametric model?
- Is a radial SVM a parametric or nonparametric model?

Bias–variance tradeoff

- What is the bias of a model?
- What is the variance of a model?
- What is the irreducible error?
- What is the sum of the bias, variance, and irreducible error, where the first two are associated with some particular error?
- What is the bias–variance tradeoff?
- What is an example of a model with low bias and high variance?
- What is an example of a model with high bias and low variance?
- How does regularizing a linear model affect its bias and variance?
- How can you distinguish models which underfit (high bias) from models which overfit (high variance)?

Numerical optimization

- What is the gradient of a function?
- What is gradient descent?
- What is stochastic gradient descent?
- What is mini-batch gradient descent?
- What is one advantage of Newton's method compared to gradient descent?
- What is one disadvantage of Newton's method compared to gradient descent?

Miscellaneous

- What is one way to combine multiple models into a single, better model?
- What is one real-world example of stacking?
- What is one way to perform a hyperparameter search aside from grid search?

- Why might you want to perform a random hyperparameter search instead of a grid search?
- What is an example of a situation when you would not want to parallelize your code with MapReduce?

Problematic datasets

Missing values

- What is a problem which might arise if you remove all rows containing missing values from your dataset?
- What is one way to fill in missing entries in a dataset?
- What is one way to use a recommender system to fill in missing entries in a dataset?

Unbalanced classes

- What is an unbalanced dataset?
- If the classes in a binary classification problem are very unbalanced, how could you correct for this?
- What is the best classification algorithm to use for an enriched classification problem?
- What classification approach would you use to detect credit card fraud?

Target leakage

- What is target leakage?
- What is one real-world example of target leakage?
- How can you deal with target leakage?

Classical statistics

- What is the mode of a set of values?
- What is the median of a set of values?
- What is the mean or expected value of a random variable?
- What is the variance of a random variable?
- What is the variance of a random expressed variable expressed as the difference of two quantities?
- What is the probability of sampling any particular specific value from a continuous probability distribution?

Theorems

- What is the law of large numbers?
- What is one example of the real-world significance of the law of large numbers?
- What is one common misconception about the law of large numbers?
- What is the central limit theorem?
- What is one example of the real-world significance of the central limit theorem?

Frequentist inference

- What is a null hypothesis?
- What is a p-value?
- What is the typical threshold for statistical significance?
- What is a t-test?
- How do a one-sample and two-sample t-test differ?
- How do a one-sided and two-sided t-test differ?
- If the p-value for a two-sided t-test is equal to C, what is the p-value for the corresponding one-sided t-test?
- If you obtain a p-value for a two-sided t-test slightly greater than 0.05, would it be appropriate to use a one-sided t-test instead to attain statistical significance?
- What is one possible explanation for a high p-value?
- What is the multiple comparisons problem?
- What is one example of a real-life situation where you might encounter the multiple comparisons problem?
- How can you correct for the multiple comparisons problem?
- What is a confidence interval?

Bayesian inference

- What is Bayes' rule?
- What is one difference between the Bayesian and frequentist interpretations of probabilities?
- What is one difference between Bayesian inference and frequentist inference?
- Is Bayesian or frequentist inference typically more computationally expensive?
- What numerical methods are used to handle Bayesian inference?

A/B testing

- What do the "A" and "B" stand for in an A/B test?
- What variables need to be taken into account to determine the length of an A/B test?
- Suppose that you run an A/B test to test an improvement to your signup button and that you run a separate A/B test for each of the 50 states. What p-value should be your threshold for statistical significance?
- What is an A/B/n test?
- Will an A/B/n test take more or less time than an A/B test?
- What is one disadvantage of using an A/B/n test compared to a series of multiple A/B tests?

Distributions

- What is a uniform distribution?
- You can choose random points on a circle's circumference. How can you use this to draw samples from a uniform distribution?
- Consider a uniform distribution from a to b from which we sample the points $x_1, x_2, \ldots, x_n$. What are the maximum likelihood estimates of a and b?
- What does the Bernoulli distribution model?
- What is one real-world example of a quantity modeled by a Bernoulli distribution?
- What is the probability density function of a Bernoulli distribution with success probability p?
- What is the expected value of a Bernoulli distribution with success probability p?
- What is the variance of a Bernoulli distribution with success probability p?
- What does the binomial distribution model?
- What is one real-world example of a quantity modeled by a binomial distribution?
- How does the binomial distribution relate to the Bernoulli distribution?
- What is the probability density function of a binomial distribution with n trials and success probability p?
- What is linearity of expectation?
- What is the expected value of a binomial distribution with n trials and success probability p?
- What is the variance of a binomial distribution with n trials and success probability p?