Review Questions

CS780/880: Introduction to Machine Learning

1 First Half-term

The following are review questions for the midterm exam. The questions are adapted from various sources (some listed below) and are the type of questions that would be asked as a part of a machine learning job interview. Some of the questions go somewhat beyond the material that we covered and those questions will not be on the midterm. The questions are in no particular order.

1. What's the trade-off between bias and variance?
2. What is the difference between supervised and unsupervised machine learning?
3. How is KNN different from k-means clustering?
4. Explain how a ROC curve works.
5. Define precision and recall.
6. What is Bayes' Theorem? How is it useful in a machine learning context?
7. Why is "Naive" Bayes naive?
8. Explain the difference between L1 and L2 regularization.
9. What's the F1 score? How would you use it?
10. What's the difference between a generative and discriminative model?
11. Which is more important to you–model accuracy, or model performance?
12. When should you use classification over regression?
13. How do you ensure you're not overfitting with a model?
14. What evaluation approaches would you work to gauge the effectiveness of a machine learning model?
15. How would you evaluate a logistic regression model?
16. How do you handle missing or corrupted data in a dataset?
17. What are your favorite use cases of machine learning models?
18. Is rotation necessary in PCA? If yes, Why? What will happen if you don't rotate the components?
19. Explain prior probability, likelihood and marginal likelihood in context of naive Bayes algorithm?
20. You came to know that your model is suffering from low bias and high variance. Which algorithm should you use to tackle it? Why?

21. You are given a data set. The data set contains many variables, some of which are highly correlated and you know about it. Would you remove correlated variables before using PCA? Why? *hint: duplicate variables?*

22. How is True Positive Rate and Recall related? Write the equation.

23. Can the (training) R-squared of a multiple linear regression model improve after removing/adding features?

24. When is Ridge regression favorable over Lasso regression?

25. What is the difference between covariance and correlation?

26. While working on a data set, how do you select important variables?

27. Rise in global average temperature led to decrease in number of pirates around the world. Does that mean that decrease in number of pirates caused the climate change?

28. You've got a data set to work having p (no. of variable) > n (no. of observation). Why is OLS as bad option to work with? Which techniques would be best to use? Why?

29. You are working on a classification problem. For validation purposes, you've randomly sampled the training data set into train and validation. You are confident that your model will work incredibly well on unseen data since your validation accuracy is high. However, you get shocked after getting poor test accuracy. What went wrong?

30. You have been asked to evaluate a regression model based on R-squared, adjusted R-squared. What will be your criteria?

31. In k-means or kNN, we use euclidean distance to calculate the distance between nearest neighbors. Why not manhattan distance ?

32. Linear regression model is generally evaluated using adjusted R-squared or F-value. How would you evaluate a logistic regression model?

33. Considering the long list of machine learning algorithm, given a data set, how do you decide which one to use?

34. Do you suggest that treating a categorical variable as continuous variable would result in a better predictive model?

35. When does regularization becomes necessary in machine learning?

36. R-squared (or RSS) is to linear regression as likelihood is to logistic regression. Explain the statement.

37. Questions about a particular model:

    (a) What type of problem does the model try to solve?

    (b) Is it prone to over-fitting? If so—what can be done about this?

    (c) Does the model make any important assumptions about the data? When might these be unrealistic? How do we examine the data to test whether these assumptions are satisfied?

    (d) Does the model have convergence problems? Does it have a random component or will the same training data always generate the same model? How do we deal with random effects in training?

    (e) What types of data (numerical, categorical etc. . .) can the model handle?

    (f) Can the model handle missing data? What could we do if we find missing fields in our data?

(g) How interpretable is the model?

(h) What alternative models might we use for the same type of problem that this one attempts to solve, and how does it compare to those?

(i) Can we update the model without retraining it from the beginning?

(j) How fast is prediction compared to other models? How fast is training compared to other models?

(k) Does the model have any meta-parameters and thus require tuning? How do we do this?

38. What is the EM algorithm? Give a couple of applications

39. What methods for dimensionality reduction do you know and how do they compare with each other?

40. How would you validate a model you created to generate a predictive model of a quantitative outcome variable using multiple regression.

41. Explain what precision and recall are. How do they relate to the ROC curve?

42. How can you prove that one improvement you've brought to an algorithm is really an improvement over not doing anything?

43. Explain what resampling methods are and why they are useful. Also explain their limitations.

44. Is it better to have too many false positives, or too many false negatives? Explain.

45. Lasso regression uses the L1-norm of coefficients as a penalty term, while ridge regression uses the L2-norm. Which of these regularization methods is more likely to result in sparse solutions?

46. Why overfitting happens?

47. What is "training set" and "test set"?

48. What is the difference between "unsupervised learning" and "supervised learning"?

49. What is classifier in machine learning?

50. What are the advantages of Naive Bayes?

51. What is K-means? How can you select K for K-means?

52. What is multicollinearity and how you can overcome it?

53. What is the curse of dimensionality?

54. How do you decide whether your linear regression model fits the data?

55. What is the difference between squared error and absolute error? Which one is better?

56. Why L1 regularizations causes parameter sparsity whereas L2 regularization does not?

57. Can you write the formula to calculate R-squared?

58. Suppose that we want to estimate the uncertainty in the prediction of a linear regression. Lets say we use bootstrapping to estimate the uncertainty (possible values) of β based on the available data. For example, we estimate that it is likely that $\beta_0 \in [1.2, 1.4], \beta_1 \in [2.3, 2.4]$. Say you are given a data point $x_i = 0$. Why it is not sufficient to expect that $y_i \in [1.2, 1.4]$ but that the variability of the target can be greater?

59. What do you understand by feature vectors?

60. How do data management procedures like missing data handling make selection bias worse?
61. What are the advantages and disadvantages of using regularization methods like Ridge Regression?
62. What do you understand by outliers and inliers? What would you do if you find them in your dataset?
63. What are the basic assumptions to be made for linear regression?
64. What do you understand by conjugate-prior with respect to Naive Bayes?
65. What is the difference between Bayesian Inference and Maximum Likelihood Estimation (MLE)?
66. What are categorical variables?
67. How can outlier values be treated?
68. How can you assess a good logistic model?
69. During analysis, how do you treat missing values?
70. Can you cite some examples where a false positive is important than a false negative?
71. Can you cite some examples where a false negative important than a false positive?
72. Can you cite some examples where both false positive and false negatives are equally important?
73. Can you explain the difference between a test set and a validation set?
74. What kind of properties are ideal for a machine learning data set?
75. Will the solution of OLS (ordinary least squares = linear regression) on log odds be the same as logistic regression? Which one is better and why?
76. What is the difference between k-means and EM for Gaussian Mixtures?
77. Pick an algorithm you like and walk me through the math and then the implementation of it, in pseudo-code.
78. Discuss how you go about feature engineering (look for both intuition and specific evaluation techniques).
79. Give us a common application of machine learning that you get to see on a daily basis?
80. How to assess the quality of clustering, especially to know when you have the right number of clusters?
81. Explain what regularization is and why it is useful
82. We know that one hot encoding increasing the dimensionality of a data set. But, label encoding does not. How?
83. What are the differences between LDA, QDA, and Naive Bayes?

2 Second Half-term

These questions are preliminary and will be edited later.

1. What cross-validation technique would you use on a time series dataset?
2. What's your favorite algorithm, and can you explain it to me in less than a minute?
3. What's the difference between Type I and Type II error?
4. What's a Fourier transform?
5. What is deep learning, and how does it contrast with other machine learning algorithms?
6. How is a decision tree pruned?
7. How would you handle an imbalanced dataset?
8. Name an example where ensemble techniques might be useful.
9. What's the "kernel trick" and how is it useful?
10. Do you have experience with Spark or big data tools for machine learning?
11. How would you implement a recommendation system for our company's users?
12. How can we use your machine learning skills to generate revenue?
13. How would you approach the "Netflix Prize" competition?
14. Where do you usually source datasets?
15. How do you think Google is training data for self-driving cars?
16. How would you simulate the approach AlphaGo took to beat Lee Sidol at Go?
17. You are given a data set. The data set has missing values which spread along 1 standard deviation from the median. What percentage of data would remain unaffected? Why?
18. You are working on a time series data set. You manager has asked you to build a high accuracy model. You start with the decision tree algorithm, since you know it works fairly well on all kinds of data. Later, you tried a time series regression model and got higher accuracy than decision tree model. Can this happen? Why?
19. You are assigned a new project which involves helping a food delivery company save more money. The problem is, company's delivery team aren't able to deliver food on time. As a result, their customers get unhappy. And, to keep them happy, they end up delivering food for free. Which machine learning algorithm can save them?
20. After spending several hours, you are now anxious to build a high accuracy model. As a result, you build 5 GBM models, thinking a boosting algorithm would do the magic. Unfortunately, neither of models could perform better than benchmark score. Finally, you decided to combine those models. Though, ensembled models are known to return high accuracy, but you are unfortunate. Where did you miss?
21. After analyzing the model, your manager has informed you that your regression model is suffering from multicollinearity. How would you check if he's true? Without losing any information, can you still build a better model?

22. Is it possible capture the correlation between continuous and categorical variable? If yes, how?

23. Both being tree based algorithm, how is random forest different from Gradient boosting algorithm (GBM)?

24. Running a binary classification tree algorithm is the easy part. Do you know how does a tree splitting takes place i.e. how does the tree decide which variable to split at the root node and succeeding nodes?

25. You've built a random forest model with 10000 trees. You got delighted after getting training error as 0.00. But, the validation error is 34.23. What is going on? Haven't you trained your model perfectly?

26. What is convex hull?

27. What cross validation technique would you use on time series data set? Is it k-fold or LOOCV?

28. You are given a data set consisting of variables having more than 30% missing values? Let's say, out of 50 variables, 8 variables have missing values higher than 30%. How will you deal with them?

29. "People who bought this, also bought. . . " recommendations seen on amazon is a result of which algorithm?

30. What is deep learning and what are some of the main characteristics that distinguish it from traditional machine learning

31. What is linear in a generalized linear model?

32. What is a probabilistic graphical model? What is the difference between Markov networks and Bayesian networks?

33. Give an example of an application of non-negative matrix factorization

34. On what type of ensemble technique is a random forest based? What particular limitation does it try to address?

35. What are some good ways for performing feature selection that do not involve exhaustive search?

36. Are you familiar with price optimization, price elasticity, inventory management, competitive intelligence? Give examples

37. What is selection bias, why is it important and how can you avoid it?

38. When training a support vector machine, what value are you optimizing for?

39. When training a 10-layer neural net using backpropagation, I find that the weights for the top 3 layers are not changing at all! The next few layers (4-6) are changing, but very slowly. What's going on and how do I fix this?

40. What is inductive machine learning?

41. What is the difference between heuristic for rule learning and heuristics for decision trees?

42. What are Bayesian Networks (BN) ?

43. What are the two classification methods that SVM (Support Vector Machine) can handle?

44. What is ensemble learning?

45. Why ensemble learning is used?

46. When to use ensemble learning?
47. What are the two paradigms of ensemble methods?
48. What is the general principle of an ensemble method and what is bagging and boosting in ensemble method?
49. What is bias-variance decomposition of classification error in ensemble method?
50. What is PCA, KPCA and ICA used for?
51. What are support vector machines?
52. What is PAC Learning?
53. What are Recommender Systems?
54. Why data cleaning plays a vital role in analysis?
55. What do you understand by the term Normal Distribution?
56. What is Collaborative filtering?
57. What does P-value signify about the statistical data?
58. Do gradient descent methods always converge to same point?
59. How you can make data normal using Box-Cox transformation?
60. What is the goal of A/B Testing?
61. What is an Eigenvalue and Eigenvector?
62. What is Gradient Descent?
63. Explain about the box cox transformation in regression models.
64. You created a predictive model of a quantitative outcome variable using multiple regressions. What are the steps you would follow to validate the model?
65. How can you deal with different types of seasonality in time series modeling?
66. In experimental design, is it necessary to do randomization? If yes, why?
67. What is the importance of having a selection bias?
68. Give some situations where you will use an SVM over a RandomForest Machine Learning algorithm and vice-versa.
69. Discuss MapReduce (or your favorite parallelization abstraction). Why is MapReduce referred to as a "shared-nothing" architecture (clearly the nodes have to share something, no?) What are the advantages/disadvantages of "shared-nothing"?
70. What are the trade-offs between closed-form and iterative implementations of an algorithm, in the context of distributed systems?
71. What is active learning?
72. How is the multi-armed bandit model different from active learning?
73. What sort of optimization problem would you be solving to train a support vector machine?

74. Tell me about positives and negatives of using Gaussian processes / general kernel methods approach to learning.

75. How does a kernel method scale with the number of instances?

76. Describe ways to overcome scaling issues with with respect to Nystrom Method, or low-rank matrix decomposition.

3 Sources

- https://www.springboard.com/blog/machine-learning-interview-questions/
- https://www.analyticsvidhya.com/blog/2016/09/40-interview-questions-asked-at-startups-in-mach
- https://resources.workable.com/machine-learning-engineer-interview-questions
- http://www.kdnuggets.com/2016/02/21-data-science-interview-questions-answers.html
- http://career.guru99.com/top-50-interview-questions-on-machine-learning/
- https://www.dezyre.com/article/100-data-science-interview-questions-and-answers-general-for-2
  184