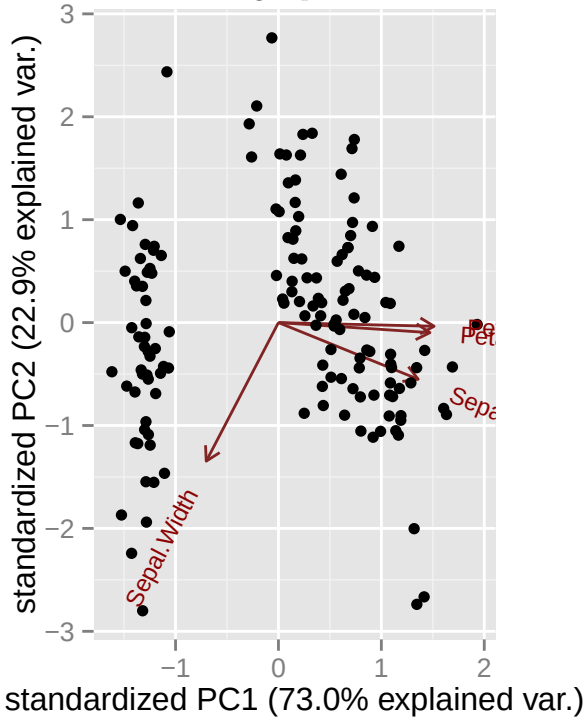
Questions

Jonathan Rosenblatt

June 1, 2015

Sample Questions

1. Based on the following Biplot...



   a. How many variables were in the original data?
   b. What original variables are captured by the first principal component?
   c. What original variables are captured by the second principal component?
   d. How many groups/clusters do you see in the data?

2.
```
n <- 100
p <- 10
X <- rnorm(n*p) %>% matrix(ncol = p, nrow=n)
sigma <- 1e1
epsilon <- rnorm(n, mean = 0, sd = sigma)
y <- X %*% beta + epsilon
```

   a. What does the code do?
   b. What is the dimension of `beta`?
   c. Can I fit a neural network to the data? Explain.

3. How does the graphical model alleviate the parameter dimensionality problem?

4. What is the difference between FA and ICA.

5. What is the cutoff of OLS classification with -1,3 encoding.

6. Name three clustering methods. Explain them.

7. You want to cluster individuals based on their LinkedIn acquaintances: name an algorithm you **cannot** use.

8.
```
hmmm <- 10
ahhh <- sample(1:5, nrow(data), replace = TRUE)
that <- NULL

for (yup in 1:hmmm){
  wow <- data[ahhh!=yup,]
  arrrg <-  data[ahhh==yup,]
  ok <- lm(y~. ,data = wow)
  nice <- predict(ok, newdata=arrrg)
  good <-  nice - arrrg$y
  that <- c(that, good)
}

MSE(that)
```

   a. What is the method implemented in the code?
   b. What problem does the method solve?

9.
```
y1 <- prcomp(.iris, scale. = TRUE)
y2 <- y1$x[,1:2]
y3 <- glm(.iris.y~y2)
```

   a. Knowing that `.iris.y` is a two-level categorical variable, what does the code do?
   b. What could be a motivation for the proposed method?

10.
```
y1 <- prcomp(.iris, scale. = TRUE)
y2 <- y1$x[,1:2]
y3 <- kmeans(y2,3)
```

   a. What does the code do?
   b. What can be the motivation for the proposed method?

11. Two scientists claim to have found two unobservable movie attributes, that drive viewers' satisfaction in the Netflix data (movie ratings data). They both used the same data and factor analysis. One claims the factors are the "action factor" and "drama factor". The other claims it is "comedy factor" and the "animation factor". Try to resolve the situation with your knowledge of factor analysis.

12. $argmin_\beta\{\frac{1}{n}\sum_i(y_i - x_i\beta)^2 + \lambda/2\|\beta\|_2^2\}$

   a. What is the name of the problem above?
   b. Does the solution enjoy the sparsity property?
   c. What is the regularization parameter? Name two methods for choosing it.

13. For the purpose of interpreting the predictor, would you prefer the CART or the NNET? Explain.

14. In order to estimate the covariance matrix in a Gaussian graphical model: should I estimate it directly or via its inverse? Explain.

15. Describe a method for selecting the number of mixing components in a mixture model using train-test samples.

16. Describe the stages of an algorithm to simulate n samples from a two-state hidden Markov model. Assume you can generate data from Bernoulli and Gaussian distributions.

17. What assumption in ICA solves the FA rotation problem?

18. What is the LASSO ERM problem? Write the formula.

19. What is the OLS ERM problem? Write the formula.

20. What is the ridge ERM problem? Write the formula.

21. Name two algorithms for unbiased estimation of the population risk $R(\theta)$.

22. Name two unbiased estimators of the in-sample–prediction-error: $\bar{R}(f) := \frac{1}{n}\sum_i E_Y[l(Y, f(x_i))]$.

23. Suggest an algorithm to choose the number of principal components using cross validation. Write in pseudo-code.

24. Can the principal components in the PCA problem be estimated using maximum likelihood? Explain.

25. What can the logistic regression estimate that the SVM cannot?

26. Can any function be approximated using the LASSO? Put differently- does the LASSO have the Universal Approximator property?

27. Write the Bernoulli likelihood loss function. To what type of y does it apply? What class of `R` objects holds this data type?

28. Name two methods for dimensionality reduction in supervised learning. Explain each briefly.

29. Here is some pseudo-code:

    - Set M candidate learning algorithms.
    - For $m \in 1, \ldots, M$, do
      - $\hat{f}^m(x) :=$ the predictor learned with the m'th algorithm.
    - EndFor
    - Set $\bar{f}(x) := \frac{1}{M}\sum_{m=1}^{M} \hat{f}^m(x)$.
    - Return $\bar{f}(x)$.

    a. What is the name of the method above?
    b. What is the problem the method is designed to solve?
    c. Suggest an improvement to the method.

30. How many parameters need to be estimated to learn a multivariate Gaussian distribution where $p = 15$. How does a graphical model help with this problem?

31.
```
##   lhs                          rhs              support confidence lift
## 1 {Instant food products,
##    soda}                  => {hamburger meat} 0.00122     0.6316   19
```

    a. What method will return this output?
    b. Interpret the output.

32. One researcher applied k-means clustering on the first two PCs. Another applied k-medoids on the output of classical MDS with Euclidean distances. Can the clusters differ? Explain.

33. Suggest a method to visualize a social network. Explain.

34. A researcher wishes to cluster songs (not the lyrics. the actual audio files). Suggest two methods that will allow this and discuss their possible advantages and disadvantages.

35. What is the difference between "complete" and "single" linkage in agglomerative clustering?

36. $(X'X + \lambda I)^{-1}X'y$. This is the solution to what problem?

37. What will happen if we try to learn an empirical risk minimizer with no inductive bias? What is the name of the phenomenon?

38. Name two justifications for the regularization term in LASSO regression. How do we know predictions can only improve with a small regularization?

39. What method learns a hypothesis in the class $f(x) = \sum_{m=1}^{M} c_m I_{\{x \in R_m\}}$.

    a. What is the name of the hypothesis class?
    b. Name a particularly desirable property of this class (and thus- of the method)

40. If I am using the Deviance likelihood as a loss function– what type is my predicted variable?

41. Having learned a mixture distribution $p(x) = \sum_{k=1}^{k} \pi_k p_k(x)$; how can I use it for clustering?

42. Why can't we produce a bi-plot for MDS while we can for PCA?

43. What is the difference between a Streaming Algorithm, and a Batch-Algorithm.

44. Why is prediction an easier task than classical statistical inference (from the Estimation course)?

45. What are the two historical motivations underlying PCA?

46. We saw that for the PCA problem, it suffice to know only the correlations between variables $X'X$. Why does it not suffice for OLS?

47. In what course did you cover methods for unsupervised learning of a parametric generative model? Name two learning methods?