University of Saskatchewan

Department of Computer Science

176 Thorvaldson Building
110 Science Place, Saskatoon, SK, S7N 5C9, Canada
Telephine: (306) 966-4886, Facimile: (306) 966-4884

CMPT 423/820

Winter 2020
Machine Learning

Assignment 4 – Solutions

Last Assignment

Date Due: April 13, 2020, 11:59pm **Total Marks: 76**

Question 1 (12 points):

Purpose: To observe the effect of missing data on classifier performance.

Competency Level: Basic

On the Assignment 4 Moodle page, you'll find a zipfile named A4Q1.ZIP, containing 5 versions of the Iris dataset.

- `iris.csv`: The original dataset
- `iris01.csv`: The original dataset, with 1% of the features deleted.
- `iris05.csv`: The original dataset, with 5% of the features deleted.
- `iris10.csv`: The original dataset, with 10% of the features deleted.
- `iris20.csv`: The original dataset, with 20% of the features deleted.

The deletions were made by randomly choosing a row and a column, and deleting the value stored there. Only feature values were deleted; not class labels. If you look at the CSV file, you'll see commas with nothing between them.

There are two problems to address. The practical problem of getting new data into the dataset, and then choice of which data to put there. For the first problem, we can look to Pandas methods. Here's a good place to start: https://jakevdp.github.io/PythonDataScienceHandbook/03.04-missing-values.html

Your main task is to explore how different approaches to dealing with missing data affects classifier performance.

1. Remove any sample (row) with missing data.
2. Fill in any missing data with a random value. Use the range (min, max) of the feature as the range for your random value.
3. Fill in the missing data with the mean value for the feature, regardless of class label.
4. Fill in the missing data with the mean value for the feature, given the class. For example, if the missing value appears on a row with class label L, then use the mean value of that feature, limiting the calculation to rows having label L.

It might be fun to add to this list, by building a model to predict missing data, or to apply Expectation-Maximization, or by trying a few different classifiers, but we don't have time for that kind of fun.

You are free to use any classifier you wish, as long as you use the same classifier for all approaches. If you choose one we haven't studied yet, that's fine, but you may need to read up about how it works so you can understand the results.

To answer this question, plot average classification accuracy (using 5-fold cross-validation, for example) against the proportion of missing data, for each of the approaches above. Also make a comment about the results, as if you were explaining them to someone who has not done this experiment. There's nothing deep here, just express in words what you discovered.

What to Hand In

- A PDF exported from a Jupyter Notebook that plots the average classification accuracy (y axis) vs the proportion of missing data (x-axis). All 4 approaches will be on the same plot. An explanation of your findings is included.

Evaluation

- 8 marks. You made a plot of classification accuracy vs the proportion of missing data for all 4 approaches.
- 4 marks. You commented on the plots, demonstrating understanding of the results.

Solution: Grading Guidelines

- 8 marks. Your plot showed the data for the 4 techniques and the 5 data files.
  - It doesn't have to be a single plot. It could be several.
  - The plots are consistent with the solution here.
- 4 marks. You discussed and interpreted the results.
  - You noted that random was bad, and class mean was good.
  - You explained why removing data worked well.
  - There was no need to reproduce the pair plots in the explanation.

Question 2 (24 points):

Purpose: To explore the use of Support Vector Machines for Classification.

Competency Level: Basic

On the Assignment 4 Moodle page, you'll find a Jupyter Notebook names A4Q2. Complete the tasks in the notebook! The data for the notebook is generated procedurally by Scikit-Learn functions.

What to Hand In

- A PDF exported from your Jupyter Notebook with the answers to the questions.

Evaluation

1. 6 marks. Kernels!
   - (2 marks) You description of the term *kernel* demonstrated understanding.
   - (2 marks) Your description of the polynomial kernel demonstrated understanding.
   - (2 marks) Your description of the Radial Basis function kernel demonstrated understanding.
2. 6 marks. Balance between error and model complexity.
   - (2 marks) Your discussion on the behaviour of the `linear` kernel as `C` changes reflects an understanding of SVM.
   - (2 marks) Your discussion on the behaviour of the `poly` kernel as `C` changes reflects an understanding of SVM.
   - (2 marks) Your discussion on the behaviour of the `rbf` kernel as `C` changes reflects an understanding of SVM.
3. 3 marks. Explore the Polynomial kernel
   - (3 marks) Your explanation of the behaviour of the polynomial kernel with different degrees reflects an understanding of SVM.
4. 3 marks. Explore the Radial Basis Function kernel
   - (3 marks) Your explanation of the behaviour of the Radial Basis Function kernel with different degrees reflects an understanding of SVM.
5. 6 marks. Other datasets
   - (3 marks) Your explanation for your choice of parameters for all three kernels reflects an understanding of the application of SVM to the `circ` dataset created by `make_circles`.
   - (3 marks) Your explanation for your choice of parameters for all three kernels reflects an understanding of the application of SVM to the `moon` dataset created by `make_moons`.

Solution: See the Jupyter Notebook `A4Q1_Solution.ipynb` or the exported document `A4Q1_Solution.pdf`.

Question 3 (12 points):

Purpose: To explore unsupervised clustering methods with KMeans and Gaussian Mixture Models.

Competency Level: Basic

On the Assignment 4 Moodle page, you'll find a Jupyter Notebook named A4Q3. Complete the tasks in the notebook! The data for the notebook is in the file `a4q3.csv`.

What to Hand In

- A PDF exported from your Jupyter Notebook with tasks completed and answers to the questions.

Evaluation

- 3 marks. Step 3.
  - You called the KMeans constructor with appropriate arguments and options.
- 3 marks. Step 4.
  - You called the GMM constructor with appropriate arguments and options.
- 3 marks. Answer to question 1
- 3 marks. Answer to question 2

Solution: See the Jupyter Notebook `A4Q3_Solution.ipynb` or the exported document `A4Q3_Solution.pdf`.

Grading Guidelines

- Step 3. 3 marks. Give full marks here unless the required options are missing.
  1. The number of clusters is required. There's no reason for it to be some value different from 5 or `n_classes`.
  2. Later parts of the notebook require changing `init` so either value could appear here. Also, the default value is useful, so the `init` might not be set explicitly.
  3. Other options might be chosen, but none are required.
- Step 4. 3 marks. Give full marks here unless the required options are missing.
  1. The number of clusters is required. There's no reason for it to be some value different from 5 or `n_components`.
  2. Later parts of the notebook require changing `init_params` so either value could appear here. Also, the default value is useful, so the `init_params` might not be set explicitly.
  3. Other options might be chosen, but none are required.
- Answer to question 1 above (3 marks).
  - For full marks, the comment should note that there is no obvious effect. No explanation is needed.
- Answer to question 2 above (3 marks).
  - For full marks, the comment should note that using random seems to put the centroids at the same location in the middle of the data. There was no requirement to explain why it happens.

Question 4 (12 points):

Purpose: To explore the use of Principal Components Analysis for dimensionality reduction

Competency Level: Basic

On the Assignment 4 Moodle page, you'll find a Jupyter Notebook named A4Q4. Complete the tasks in the notebook! The data for the notebook is in the file `a4q4.csv`.

What to Hand In

- A PDF exported from your Jupyter Notebook with tasks completed and answers to the questions.

Evaluation

- 3 marks: You built a classifier using the whole dataset, and reported an accuracy value.
- 3 marks: You applied PCA, and visualized the first two principle components.
- 3 marks: You built a classifier using the first two principle components, and reported an accuracy value.
- 3 marks: You reported on the difference in accuracy between your two classifiers.

Solution: See the Jupyter Notebook `A4Q4_Solution.ipynb` or the exported document `A4Q4_Solution.pdf`.

Grading Guidelines

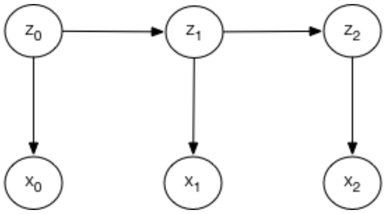
1. 3 marks: You built a classifier using the whole dataset, and reported an accuracy value.
   - Full marks for any classifier and any accuracy results.
     - Naive Bayes: around 90% accuracy.
     - SVM (linear): around 90-100% accuracy.
     - KNN (k=11): around 100% accuracy.
2. 3 marks: You applied PCA, and visualized the first two principle components.
   - Full marks for calling PCA, and plotting the results.
     - Nice neat clusters!
3. 3 marks: You built a classifier using the first two principle components, and reported an accuracy value.
   - Full marks for any classifier and any accuracy results.
     - Naive Bayes: around 100% accuracy.
     - SVM (linear): around 90-100% accuracy.
     - KNN (k=11): around 100% accuracy.
4. 3 marks: You reported on the difference in accuracy between your two classifiers.
   - Full marks for a report that compared the two results, and made some attempt to interpret the results.
     - Naive Bayes: around 100% accuracy. This was a little better than the 6D data. This means that there are useful distinctions in the 6D data that are fairly effectively used by Naive Bayes. The problem is that we can't see them in 6D.
     - SVM (linear): around 90-100% accuracy. The accuracy didn't improve, but it also didn't get worse. That means for this data we could base the classifier on the transformed data, meaning the classifier would be more efficient, and we would not lose any accuracy.
     - KNN (k=11): around 100% accuracy. Using PCA represents an opportunity to build a simpler model using far fewer features, just as described for SVC.

Question 5 (10 points):

Purpose: To clarify the HMM equations on a very small example. Don't refer to the HMM formulae in your notes for this. Just use the notation and probability basics you learned in A3. We'll see how to derive those formulae, so that they are less abstract.

Competency Level: Extended

The following diagram represents a simple, 3 step HMM represented as a 6 node Bayesian Network.



The Z_i are the state variables; the X_i are the observable "evidence." The state transition model is the CPD $P(Z_i|Z_{i-1})$. The sensor model or emission probabilities are the CPD $P(X_i|Z_i)$. The initial state is described by the CPD $P(Z_0)$. You don't have numeric probabilities, so just do the algebra, as in the previous question.

Hint: use the notions of relevance and independence developed for Bayesian networks.

(a) Show that

$$P(Z_0|X_0) = \frac{P(X_0|Z_0)P(Z_0)}{P(X_0)}.$$

Note: This is a simple form of the state estimation problem: given one observation, predict state Z_0.

Solution: This is simply an application of Bayes Theorem, but it can also be derived through the use of Variable Elimination.

$$\begin{aligned} P(Z_0|X_0) &= \frac{P(Z_0, X_0)}{P(X_0)} \\ &= \frac{P(X_0|Z_0)P(Z_0)}{P(X_0)} \end{aligned}$$

(b) Show that

$$P(X_0) = \sum_{Z_0} P(X_0|Z_0)P(Z_0).$$

Note: This is the denominator of the previous derivation. Not particularly interesting, because it is a recapitulation of the numerator. See it there?

Solution:

$$\begin{aligned} P(X_0) &= \sum_{Z_0} P(X_0, Z_0) \\ &= \sum_{Z_0} P(X_0|Z_0)P(Z_0) \end{aligned}$$

(c) Show that

$$P(Z_1|X_1,X_0) = \frac{P(X_1|Z_1)P(X_0)\sum_{Z_0} P(Z_1|Z_0)P(Z_0|X_0)}{P(X_1,X_0)}.$$

Note: This is a simple form of the state estimation problem. Given two observations, predict the state Z_1. Especially, note the factor $P(Z_0|X_0)$ within the formula. We already calculated that, above!

Solution: We can follow the Variable Elimination procedure here, with the only relevant variable being Z_0.

$$\begin{aligned}
P(Z_1|X_1X_0) &= \frac{P(Z_1X_1X_0)}{P(X_1X_0)} \\
&= \frac{\sum_{Z_0} P(Z_1Z_0X_1X_0)}{P(X_1X_0)} \\
&= \frac{\sum_{Z_0} P(X_1|Z_1)P(Z_1Z_0X_0)}{P(X_1X_0)} \\
&= \frac{P(X_1|Z_1)\sum_{Z_0} P(Z_1|Z_0)P(Z_0X_0)}{P(X_1X_0)} \\
&= \frac{P(X_1|Z_1)\sum_{Z_0} P(Z_1|Z_0)P(Z_0|X_0)P(X_0)}{P(X_1X_0)} \\
&= \frac{P(X_1|Z_1)P(X_0)\sum_{Z_0} P(Z_1|Z_0)P(Z_0|X_0)}{P(X_1X_0)}
\end{aligned}$$

(d) Show that

$$P(X_0,X_1) = P(X_0)\sum_{Z_1} P(X_1|Z_1)\sum_{Z_0} P(Z_1|Z_0)P(Z_0|X_0).$$

Note: This is the denominator of the previous derivation. Not particularly interesting, because it is a recapitulation of the numerator.

Solution: This is the denominator of the previous derivation. We could try to use variable elimination, but it's better to see the denominator as a marginalization of the numerator from the previous derivation, since that's why we need it!

$$\begin{aligned}
P(X_0,X_1) &= \sum_{Z_1} P(X_1|Z_1)P(X_0)\sum_{Z_0} P(Z_1|Z_0)P(Z_0|X_0) \\
&= P(X_0)\sum_{Z_1} P(X_1|Z_1)\sum_{Z_0} P(Z_1|Z_0)P(Z_0|X_0)
\end{aligned}$$

Clarification: Notice that the factor $P(X_0)$ appears in the numerator (Q4.c) and in the denominator (Q4.d). This factor is constant relative to the query variable Z of Q4.c. Mathematically it cancels (assuming that we're not using impossible observations as evidence). As a result, we might give the formula for Q4.c as follows:

$$P(Z_1|X_1,X_0) = \alpha P(X_1|Z_1)\sum_{Z_0} P(Z_1|Z_0)P(Z_0|X_0).$$

(e) Show that

$$P(Z_2|X_2,X_1,X_0) = \frac{P(X_2|Z_2)P(X_1X_0)\sum_{Z_1} P(Z_2|Z_1)P(Z_1|X_1,X_0)}{P(X_2,X_1,X_0)}.$$

Note: You've done this 3 times now, deriving basically the same formula. This is the forward phase of the forward-backward algorithm for HMMs, expressed recursively: $P(Z_2|X_2, X_1, X_0)$ is expressed in terms of $P(Z_1|X_1, X_0)$. The steps (a) through (e) reveal how the forward algorithm for HMMs works.

Solution: The key to this derivation is that we should not use Variable elimination. While it will result in a correct formulae, it does not result in a formulae that expresses the impact of the evidence on state Z_2 in terms of the evidence on the impact of a previous state, Z_1. For that reason, we will use Z_1 only in our summation:

$$P(Z_2|X_2, X_1, X_0) = \frac{P(Z_2, X_2, X_1, X_0)}{P(X_2, X_1, X_0)}$$

Now let's work with the numerator, so that the clutter is diminished. First we'll use the ancestor Z_1, and then we'll factorize, starting with X_2.

$$\begin{aligned} P(Z_2, X_2, X_1, X_0) &= \sum_{Z_1} P(Z_2, Z_1, X_2, X_1, X_0) \\ &= \sum_{Z_1} P(X_2|Z_2, Z_1, X_1, X_0) P(Z_2, Z_1, X_1, X_0) \end{aligned}$$

By the rules of Bayes Ball, X_2 is independent of every other variable in the expression, given Z_2:

$$\begin{aligned} P(Z_2, X_2, X_1, X_0) &= \sum_{Z_1} P(X_2|Z_2) P(Z_2, Z_1, X_1, X_0) \\ &= P(X_2|Z_2) \sum_{Z_1} P(Z_2, Z_1, X_1, X_0) \end{aligned}$$

Now we work on the second factor:

$$\begin{aligned} P(Z_2, X_2, X_1, X_0) &= P(X_2|Z_2) \sum_{Z_1} P(Z_2, Z_1, X_1, X_0) \\ &= P(X_2|Z_2) \sum_{Z_1} P(Z_2|Z_1, X_1, X_0) P(Z_1, X_1, X_0) \\ &= P(X_2|Z_2) \sum_{Z_1} P(Z_2|Z_1) P(Z_1, X_1, X_0) \end{aligned}$$

The last step above is also due to conditional independence, as revealed by the graph. At last, we work on the final term:

$$\begin{aligned} P(Z_2, X_2, X_1, X_0) &= P(X_2|Z_2) \sum_{Z_1} P(Z_2|Z_1) P(Z_1, X_1, X_0) \\ &= P(X_2|Z_2) \sum_{Z_1} P(Z_2|Z_1) P(Z_1|X_1, X_0) P(X_1, X_0) \\ &= P(X_2|Z_2) P(X_1, X_0) \sum_{Z_1} P(Z_2|Z_1) P(Z_1|X_1, X_0) \end{aligned}$$

Combining the numerator and the denominator that we left behind above:

$$\begin{aligned} P(Z_2|X_2, X_1, X_0) &= \frac{P(Z_2, X_2, X_1, X_0)}{P(X_2, X_1, X_0)} \\ &= \frac{P(X_2|Z_2) P(X_1, X_0) \sum_{Z_1} P(Z_2|Z_1) P(Z_1|X_1, X_0)}{P(X_2, X_1, X_0)} \end{aligned}$$

Clarification: The factor $P(X_1 X_0)$ in the numerator is constant relative to the query variable Z_2. Because the denominator is derived from the expression in the numerator, it will appear there too. Mathematically it cancels (assuming that we're not using impossible observations as evidence). As a result, we might rewrite the formula for Q4.e as follows:

$$P(Z_2|X_2, X_1, X_0) = \alpha P(X_2|Z_2) \sum_{Z_1} P(Z_2|Z_1)P(Z_1|X_1, X_0)$$

where α represents the quantity $\frac{P(X_1,X_0)}{P(X_2,X_1,X_0)}$. It's sole purpose is to ensure that the numerator is normalized, and we don't even need a formula to calculate it. When we have the numerator calculated for every possible value of Z_2, we add up all those values, and divide each numerator by that sum. Hence, α is called a normalization factor, and usually no formula for it is given.

What to Hand In

- Five calculations. Add these answers to a document for questions 5-6, named `A4.PDF`. You can use Jupyter notebooks, or if you prefer, a LaTeX document. If necessary, you can submit a scanned document.

Evaluation

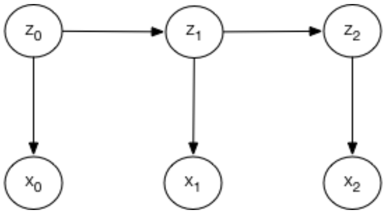
- 2 marks each.

Question 6 (6 points):

Purpose: To clarify the HMM equations on a very small example. Just use the notation and probability basics you've been practicing so far.

Competency Level: Advanced

The following diagram represents a simple, 3 step HMM represented as a 6 node Bayesian Network, the same one as before.



The Z_i are the state variables; the X_i are the observable "evidence." The state transition model is the CPD $P(Z_i|Z_{i-1})$. The sensor model or emission probabilities are the CPD $P(X_i|Z_i)$. The initial state is described by the CPD $P(Z_0)$. You don't have any numeric probabilities, so just work on the algebra, as in the previous question.

(a) Show that

$$P(X_2|Z_1) = \sum_{Z_2} P(X_2|Z_2)P(Z_2|Z_1)$$

Hint: Use conditional independence given Z_2. Don't sum over all nuisance variables!

Solution: Again, we could use Variable Elimination (Assignment 3), but it's easier just to sum over Z_2 directly:

$$\begin{aligned} P(X_2|Z_1) &= \sum_{Z_2} P(X_2, Z_2|Z_1) \\ &= \sum_{Z_2} P(X_2|Z_2, Z_1)P(Z_2|Z_1) \\ &= \sum_{Z_2} P(X_2|Z_2)P(Z_2|Z_1) \end{aligned}$$

The last step is true because when we use the rules of Bayes Ball on the graph, we realize X_2 is independent of Z_1 given Z_2.

(b) Show that

$$P(X_2, X_1|Z_0) = \sum_{Z_1} P(X_2|Z_1)P(X_1|Z_1)P(Z_1|Z_0)$$

Hint: Use conditional independence given Z_1.

Solution: We'll use the previous part as a guide. We know that given Z_1, X_2 is conditionally independent of all the other X_1, X_0. So we'll use that first:

$$\begin{aligned} P(X_2, X_1|Z_0) &= \sum_{Z_1} P(X_2, X_1, Z_1|Z_0) \\ &= \sum_{Z_1} P(X_2|Z_1, X_1, Z_0)P(Z_1, X_1|Z_0) \end{aligned}$$

$$= \sum_{Z_1} P(X_2|Z_1)P(Z_1, X_1|Z_0)$$

Now we can use the conditional independence of Z_0 and X_1, given Z_1:

$$\begin{aligned} P(X_2, X_1|Z_0) &= \sum_{Z_1} P(X_2|Z_1)P(Z_1, X_1|Z_0) \\ &= \sum_{Z_1} P(X_2|Z_1)P(X_1|Z_1, Z_0)P(Z_1|Z_0) \\ &= \sum_{Z_1} P(X_2|Z_1)P(X_1|Z_1)P(Z_1|Z_0) \end{aligned}$$

We've done all this work, leaving Z_0 "behind" the conditioning bar.
It is legitimate to begin this way:

$$P(X_2, X_1|Z_0) = \frac{P(X_2, X_1, Z_0)}{P(Z_0)}$$

And if we do, we'll eventually get a factor of $P(Z_0)$ in both the numerator and denominator, so they'll cancel. The fact that they cancel validates the choice to leave Z_0 behind the conditioning bar through the earlier derivation.

Note: This might be the first tricky derivation so far.

(c) Show that

$$P(Z_1|X_0, X_1, X_2) = \frac{P(X_2|Z_1)P(X_1, X_0)P(Z_1|X_1, X_0)}{P(X_0, X_1, X_2)}$$

Hint: Use conditional independence given Z_1.

Solution: We'll begin with the standard step of expressing a conditional as a ratio:

$$P(Z_1|X_0, X_1, X_2) = \frac{P(Z_1, X_0, X_1, X_2)}{P(X_0, X_1, X_2)}$$

To reduce clutter, we'll focus on the numerator only, and we'll start by re-using the fact that X_2 is conditionally independent of all the other X_i given Z_1:

$$\begin{aligned} P(Z_1, X_0, X_1, X_2) &= P(X_2|Z_1, X_0, X_1)P(Z_1, X_0, X_1) \\ &= P(X_2|Z_1, X_0, X_1)P(Z_1, X_0, X_1) \\ &= P(X_2|Z_1)P(Z_1, X_0, X_1) \\ &= P(X_2|Z_1)P(Z_1|X_0, X_1)P(X_0, X_1) \end{aligned}$$

Combining the numerator with the denominator (above) gives us the result.

Note: The factor $P(X_2|Z_1)$ was derived directly above, and the factor $P(Z_1|X_1, X_0)$ earlier in the assignment. This formula tells us how all the data affects the state at a single time point. Observable evidence X_2 affects Z_1 from one direction using the backward algorithm; observable evidence X_0, X_1 affects Z_1 from the other direction using the forward algorithm. To see this more clearly, we have to use a larger HMM, but the derivations are essentially the same. See the next question.

Clarification: As before, the factor $P(X_1X_0)$ in the numerator is constant relative to the query variable

Z_2. As a result, we might rewrite the formula for Q5.c as follows:

$$P(Z_1|X_0, X_1, X_2) = \alpha P(X_2|Z_1)P(Z_1|X_1, X_0)$$

What to Hand In

- Three small calculations. Add these answers to a document for questions 5-6, named `A4.PDF`. You can use Jupyter notebooks, or if you prefer, a LaTeX document. If necessary, you can submit a scanned document.

Evaluation

- 2 marks each.