CS 491 Final Exam Name Solution.

Short Problems (20 points)

1. Learning to drive is an example of imitation learning or reinforcement learning? Explain. (5 points)

Reinforcement learning. You are receiving feedback from the environment (including the instructor) and making adjustments accordingly.

2. Labels would never be helpful when using PCA. True or False? Choose one and explain. (5 points)

You could use labels to determine which PCs on which to project the data.

3. There are some situations in which a Perceptron will give better results than the SVM. True or False? Choose one and explain. (5 points)

If the data is nonlinearly separable and we use a hard-margin SVM, then SVM will say "no way, no solution" whereas we can set max-epoch with Perceptron & get an answer.

4. Consider the following situation: Landry is concerned about the security of their bank accounts. They call the bank and the manager asks them to detail all instances in the last six months wherein they might have shared their account details with another person for any kind of transaction, or may have accessed their online account from a public system, etc... Landry had ten such instances in reality, and they narrated twenty instances to finally spell out the ten correct instances. What would be Landry's recall and precision in this situation? (5 points)

recall: all 10 found → 100%

precision: 10 found out of 20 returned → 50%

Models and Algorithms (20 points)

5. For each of the methods listed below. Detail the model and the algorithm ("iterative" is not enough, write out the algorithm in steps) associated with the method. Clearly label each.

(a) 3-NN (5 points)

model: none.

algorithm: test point X_t
1. Calc distance from X_t to all train pts.
2. sort train pts. by distance.
3. Choose 3 closest pts & do majority vote.

(b) 5-Means (5 points)

model: 5 cluster centers.

algorithm:
- init. cluster centers.
- assign each training sample to nearest cluster center
- recompute cluster centers (average)

Start again until convergence.

(c) Perceptron (5 points)

model: w & b.

algorithm: $w=0$ $b=0$
check if $ya \geq 0$ (correct prediction)
if not update w & b.
go until no more updates to w & b.

(d) Decision Trees (5 points)

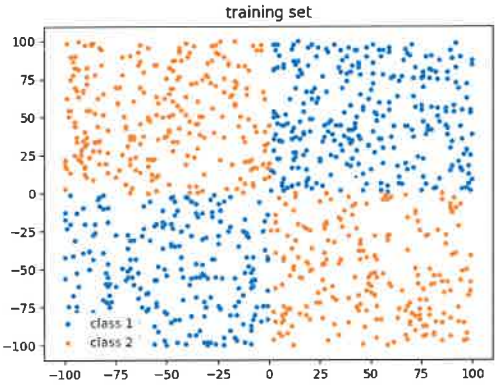
model: Binary tree

algorithm: for each feature calculate info gain if we were to split on that feature. Choose feature w/ highest info gain.
Do this recursively until max depth.

Unsupervised Pre-Processing (10 points)

6. I have the following data. I want to use an unsupervised learning method to structure the data prior to using a supervised method for classification. Choose an unsupervised learning method and a supervised learning method for your implementation. Detail the steps required to implement your combination (supervised + unsupervised) method. Justify your choice of both unsupervised and supervised methods. Assume the reader knows how the methods work, so specify hyperparameters, and what inputs and outputs would be and how you would use them.

reword.



Unsupervised: Kernels $\phi(x_1, x_2) = \{x_1, x_2\}$

This is just XOR.

~~Blue~~ orange + blue

Any supervised algorithm would work here. We can use SVM to give max margin, but a simple decision tree that uses a threshold would work nicely.

SVM And Kernels (15 points)

general case X_d.

7. Show that the following functions are kernels. That is, find the feature expansion ϕ such that $K(x, z) = \phi(x) \bullet \phi(z)$.

(a) $K(x, z) = 1 + x \bullet z$

$1 + X_1 Z_1 + X_2 Z_2 + \cdots$

$\phi(x) = \{1, X_1, X_2, X_3 \cdots, X_d\}$

(b) $K(x, z) = (x \bullet z)^2$

$(X_1 Z_1 + X_2 Z_2 + \cdots X_d Z_d)$

$X_1^2 Z_1^2 + X_2^2 Z_2^2 + \cdots + 2 X_1 Z_1 X_2 Z_2 + \cdots$

$\phi(x) = \{X_1^2, X_2^2, \cdots, \sqrt{2} X_1 X_2, \sqrt{2} X_2 X_3, \cdots\}$

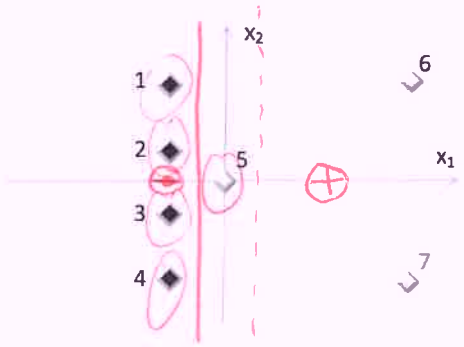
(c) $K(x, z) = (1 + x \bullet z)^2$

Same as before + $1 + X_1 Z_1 + X_2 Z_2 + \cdots$

$\phi(x) = \{1, X_1, X_2, \cdots, \sqrt{2} X_1 X_2, \cdots, X_1^2\}$

Decision Boundaries (15 points)

8. Consider a 2-class classification problem in a 2D feature space with labels (target values) that are either $+1$ or -1. The training data consists of 7 samples as shown below. (4 black diamonds for the positive class and 3 white diamonds for the negative class). The points are marked 1-7 for your convenience.



(a) K-Means: Draw on the plot the centers for the two classes. Mark them with a $\oplus$ for the positive class and $\ominus$ for the negative class. Draw the decision boundary for k-means using a dashed line. (3 points)

i. What is the training error? (1 point) $\frac{1}{7}$

ii. Is there any sample such that upon its removal, the decision boundary changes in a manner that the removed sample goes to the other side? Answer ("yes" or "no"). (1 point)

no explanation

(b) SVM: Draw the decision boundary for the hard-margin SVM using a thick solid line. Draw the margins on either side with thinner solid lines. Circle the support vectors. (3 points)

i. What is the training error? (1 point) 0

ii. The removal of which sample would change the decision boundary? Write "None" if none. (1 point) 5.

(c) Which method, K-Means of SVM, is more generalizable in this setting? Explain. (2 points)

K means. Larger margin

(d) Is there a setting in which K-Means and SVM could give the same decision boundary on this data? Yes or no. Explain. (3 points)

Soft margin SVM. dependhgon C value weighting slack.

Missing Data (10 points)

9. I have the following snippet (small portion) of a dataset containing medical data. We want to use this dataset to predict cervical cancer. There is at least one sample with a feature value missing (NaN) (there may be more, remember this is a small portion of the dataset). We learned about 3 unsupervised learning methods in this class. Choose one method and detail the steps you would use to "fill in" the missing feature values. Specify inputs and outputs to the unsupervised learning algorithm and how you would use them to replace the NaN values.

	Age	Number of sexual partners	First sexual intercourse	Num of pregnancies	Smokes	Smokes (years)	Smokes (packs/year)	Hormonal Contraceptives	Hormonal Contraceptives (years)	IUD
0	18	4.0	15.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
1	15	1.0	14.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
2	34	1.0	NaN	1.0	0.0	0.0	0.0	0.0	0.0	0.0
3	52	5.0	16.0	4.0	1.0	37.0	37.0	1.0	3.0	0.0
4	46	3.0	21.0	4.0	0.0	0.0	0.0	1.0	15.0	0.0

1. remove features w/ NaN values.
2. use k means to cluster the data.
3. take the Nan value for a feature to be the average value of the feature for the samples in the same cluster.

ML in the Real World (10 points)

10. You have developed a wonderful new theory of classification learning, called "the UNR algorithm". You input data, it outputs a formula. You decide to test your theory on the problem of predicting in January who will win a political election in November given attributes like office, region, party, incumbency, success in raising funds to date, current poll numbers, etc. You have a database with all this data for all American elections over the last twenty years. You ask a research assistant to test the UNR algorithm on this data. They come back with the good news that when they ran the UNR algorithm over the your data set, it output a formula that gives 80% accuracy on the data. What further tests should you run before release this method, claiming that you have promising new technique for machine learning? List at least three things you would test, how you would test them, and how they would provide you with more information about the UNR algorithm.

1. F-measure. Accuracy doesn't say much. F1 gives balance of Precision & Recall.

2. Cross-Validation. Were they testing on the training set?

3. Test on another problem Scenario?

CS 691 Final Exam Name Solution

Short Problems (20 points)

1. Learning to drive is an example of imitation learning or reinforcement learning? Explain. (5 points)

Same as 491 #1.

2. Labels would never be helpful when using PCA. True or False? Choose one and explain. (5 points)

Same as 491 #2.

3. Why do we convert the SVM problem to its dual form? (5 points) give 2 reasons.

1) So we can solve the problem using GD.
2) So we can use the kernel trick.

4. Consider the following situation: Landry is concerned about the security of their bank accounts. They call the bank and the manager asks them to detail all instances in the last six months wherein they might have shared their account details with another person for any kind of transaction, or may have accessed their online account from a public system, etc... Landry had ten such instances in reality, and they narrated twenty instances to finally spell out the ten correct instances. What would be Landry's recall and precision in this situation? (5 points)

Same as 491 #4

Models and Algorithms (20 points)

5. For each of the methods listed below. Detail the model and the algorithm ("iterative" is not enough, write out the algorithm in steps) associated with the method. Clearly label each.

   (a) 3-NN (5 points)

Same as 491

   (b) 5-Means (5 points)

   (c) Perceptron (5 points)

   (d) PCA (5 points)

model: k eigenvectors.

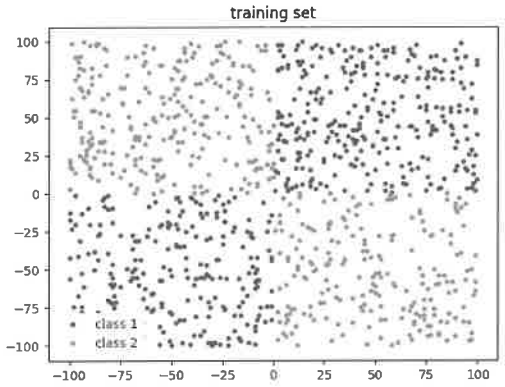
algorithm: 1) standardize data

2) find eigenvectors of covariance matrix.

3) Project data onto new space. reducing dim.

Unsupervised Pre-Processing (10 points)

6. I have the following data. I want to use an unsupervised learning method to structure the data prior to using a supervised method for classification. Choose an unsupervised learning method and a supervised learning method for your implementation. Detail the steps required to implement your combination (supervised + unsupervised) method. Justify your choice of both unsupervised and supervised methods. Assume the reader knows how the methods work, so specify hyperparameters, and what inputs and outputs would be and how you would use them.



#6

Same as #1

SVM And Kernels (15 points)

7. Multiple kernels can be combined to produce new kernel functions. For example $K(x,z) = K_1(x,z) + K_2(x,z)$ is a valid kernel function. For the questions below, kernel K_1 has the associated feature transform ϕ_1 and similarly K_2 has the feature transform ϕ_2. Identify the feature transform associated with K for the expressions given below. *Note:* The operator $[*,*]$ denotes concatenation of the two arguments. For example, $[x,z] = (x1, x2, z1, z2)$.

   (a) $K(x,z) = aK_1(x,z)$, for some scalar $a > 0$ (5 points)

   $\phi = \sqrt{a}\,\phi_1$

   (b) $K(x,z) = aK_1(x,z) + bK_2(x,z)$, for some scalars $a, b > 0$ (5 points)

   $\sqrt{a}\phi_1 \cdot \sqrt{a}\phi_1 + \sqrt{b}\phi_2 \cdot \sqrt{b}\phi_2$

   $\sqrt{a}\sqrt{b}(\phi_1 + \phi_2)$

   $[\sqrt{a}\phi_1, \sqrt{b}\phi_2]$

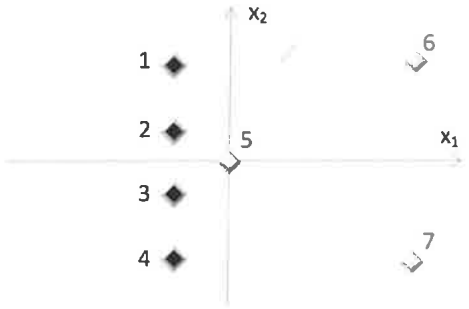
   (c) Suppose you are given the choice between using the normal perceptron algorithm, which directly works with $\phi(x)$, and the dual (kernelized) perceptron algorithm, which does not explictly compute $\phi(x)$ but instead works with the kernel function K. Keeping space and time complexities in mind, when would you prefer using the kernelized perceptron algorithm over the normal perceptron algorithm? Choose one and explain. *Note:* Here N denotes the total number of training samples and d is the dimensionality of $\phi(x)$. (5 points)

   i. $d >> N$ (d is much greater than N)
   ii. $d << N$ (d is much less than N)
   iii. Always
   iv. Never

in perceptron we replace
w w/ α & so
$|w| = d$ & $|\alpha| = N$
so it only makes sense
to use the kernel if
$|w| >> |\alpha|$

Decision Boundaries (15 points)

8. Consider a 2-class classification problem in a 2D feature space with labels (target values) that are either $+1$ or -1. The training data consists of 7 samples as shown below. (4 black diamonds for the positive class and 3 white diamonds for the negative class). The points are marked 1-7 for your convenience.



(a) K-Means: Draw on the plot the centers for the two classes. Mark them with a $\oplus$ for the positive class and $\ominus$ for the negative class. Draw the decision boundary for k-means using a dashed line. (3 points)

i. What is the training error? (1 point)

Same as #1 #8

ii. Is there any sample such that upon its removal, the decision boundary changes in a manner that the removed sample goes to the other side? Answer ("yes" or "no"). (1 point)

(b) SVM: Draw the decision boundary for the hard-margin SVM using a thick solid line. Draw the margins on either side with thinner solid lines. Circle the support vectors. (3 points)

i. What is the training error? (1 point)

ii. The removal of which sample would change the decision boundary? Write "None" if none. (1 point)

(c) Which method, K-Means of SVM, is more generalizable in this setting? Explain. (2 points)

(d) Is there a setting in which K-Means and SVM could give the same decision boundary on this data? Yes or no. Explain. (3 points)

Missing Data (10 points)

9. I have the following snippet (small portion) of a dataset containing medical data. We want to use this dataset to predict cervical cancer. There is at least one sample with a feature value missing (NaN) (there may be more, remember this is a small portion of the dataset). We learned about 3 unsupervised learning methods in this class. Choose one method and detail the steps you would use to "fill in" the missing feature values. Specify inputs and outputs to the unsupervised learning algorithm and how you would use them to replace the NaN values.

	Age	Number of sexual partners	First sexual intercourse	Num of pregnancies	Smokes	Smokes (years)	Smokes (packs/year)	Hormonal Contraceptives	Hormonal Contraceptives (years)	IUD
0	18	4.0	15.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
1	15	1.0	14.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
2	34	1.0	NaN	1.0	0.0	0.0	0.0	0.0	0.0	0.0
3	52	5.0	16.0	4.0	1.0	37.0	37.0	1.0	3.0	0.0
4	46	3.0	21.0	4.0	0.0	0.0	0.0	1.0	15.0	0.0

Same as 491 #9

ML in the Real World (10 points)

10. You have developed a wonderful new theory of classification learning, called "the UNR algorithm". You input data, it outputs a formula. You decide to test your theory on the problem of predicting in January who will win a political election in November given attributes like office, region, party, incumbency, success in raising funds to date, current poll numbers, etc. You have a database with all this data for all American elections over the last twenty years. You ask a research assistant to test the UNR algorithm on this data. They come back with the good news that when they ran the UNR algorithm over the your data set, it output a formula that gives 80% accuracy on the data. What further tests should you run before release this method, claiming that you have promising new technique for machine learning? List at least three things you would test, how you would test them, and how they would provide you with more information about the UNR algorithm.

Same as #9 #10.