Unit 3 - Probability Theory

Quiz #1

The due date for submitting this assignment has passed. **Due on 2017-02-08, 23:59 IST.**

Assignment submitted on 2017-02-07, 23:43 IST

- This quiz consists of 10 questions on 'Introduction to Machine Learning' and 'Probability'.
- All questions are mandatory and will count towards your assessment/grades.
- The following is the break-up of the first 10 questions.
  - Questions 1 to 4 carry 1 point each
  - Questions 5 to 8 carry 2 points each
  - Questions 9 and 10 carry 3 points each
- Any doubts with regards to this quiz shall be restricted to a dedicated discussion thread for this quiz.

All the best!

1) A new phone, E-Corp X1 has been announced and it is what you've been waiting for, all along. You decide to read the reviews before buying it. From past experiences, you've figured out that good reviews mean that the product is good 90% of the time and bad reviews mean that it is bad 70% of the time. Upon glancing through the reviews section, you find out that the X1 has been reviewed 1269 times and only 127 of them were bad reviews. What is the probability that, if you order the X1, it is a bad phone? ***1 point***

- 0.1362
- (•) 0.160
- 0.840
- 0.773

Yes, the answer is correct.
Score: 1

Accepted Answers:
0.160

2) Find the **least** appropriate combination of a problem and the type of method used to solve it. ***1 point***

- Predicting rainfall based on environmental indicators - **Regression**
- Predicting if Christopher Nolan will release a film this month - **Classification/Discrimination**
- (•) Finding the type of scenery in the background of a photograph - **Regression**
- Finding the optimal current gear to use in an automatic-transmission vehicle based on external indicators - **Regression**

Yes, the answer is correct.
Score: 1

Accepted Answers:
*Finding the type of scenery in the background of a photograph - **Regression***

3) A DTH TV connection uses a satellite dish fixed on a window sill with no roof for protection. When it's raining, 3 out the 20 times you observe that the TV works in spite of the rain. Out of a total of 200 times that you've used the TV, you find that it hasn't worked a total of 27 times for whatever reason.
If you turn on the TV and it doesn't work, what is the probability that it is raining? ***1 point***

- 0.017
- 0.725
- (•) 0.627
- 0.93

 is correct.

Score: 1

Accepted Answers:
0.627

4) Select the curve complexity that is most appropriate for the given dataset. ***1 point***

- ○ Linear
- ◉ Quadratic
- ○ Cubic

Yes, the answer is correct.
Score: 1

Accepted Answers:
Quadratic

5) You have two people, A and B. You place integrity over everything and hence, you don't allow any form of communication during problem solving and you've made that clear to both A and B. When placed in different rooms, A solved 25 problems out of 30 while B solved 37 out of 40. You now decided to place A and B in the same room. They seemed to solve 14 out of 16 problems. You now tell the participants that you'll be observing them on camera while they solve problems(although you can afford no such provisions!). In different rooms, A solves 12 of 15 questions, while B Solves 16 of 20 questions. In the same room, they solve 32 out of 50 questions.
Based on the above information, can you conclude if A and B were dishonest when they thought no one was watching? Were they dishonest even when they thought they were being watched? ***2 points***

- ○ Yes and Yes. A and B were dishonest and they should be reprimanded.
- ◉ No and No. A and B are model students, show integrity even when faced with difficult choices and should be rewarded.
- ○ Yes and No. This kind of behaviour is not acceptable and they should be reprimanded.
- ○ No and Yes. For some strange reason, they seem to get a thrill out of showing dishonesty when being watched. You think there might be something wrong with them.

No, the answer is incorrect.
Score: 0

Accepted Answers:
Yes and No. This kind of behaviour is not acceptable and they should be reprimanded.

6) One of the most common uses of Machine Learning today is in the domain of Robotics. Robotic tasks include a multitude of ML methods tailored towards navigation, robotic control and a number of other tasks. Robotic control includes controlling the actuators available to the robotic system. An example of this is control of a painting arm in automotive industries. The robotic arm must be able to paint every corner in the automotive parts while minimizing the quantity of paint wasted in the process. Which of the following learning paradigms would you select for training such a robotic arm? ***2 points***

- ○ Supervised Learning
- ○ Unsupervised Learning
- ◉ Reinforcement Learning

Yes, the answer is correct.
Score: 2

Accepted Answers:
Reinforcement Learning

7) Before applying a Machine Learning model, one typically analyzes the data and its properties to gain valuable insights before proceeding to learn a model over the same. One of the most common forms of analysis involves checking for correlation between input variables/fields. If the input data contains *K* **features**, the result of such an analysis is a *KxK* **covariance** matrix, Σ where $\Sigma_{i,j} = Cov(X_i, X_j)$ and X_i is the i^{th} feature. Essentially, the covariance matrix *summarizes* the data in it's own way. ***2 points***

If you're provided with some data involving *K* input features, and the *actual* distribution, from which the points are drawn, exhibit the following property, for every data point $\langle x_1, x_2 \ldots x_K \rangle$, what can you conclude about the covariance matrix?
$p(\langle X_1, X_2 \ldots X_K \rangle = \langle x_1, x_2 \ldots x_K \rangle) = \Pi_{i=1}^{i=K} p(X_i = x_i)$
Note that $1 \leq i, j \leq K \; \forall i, j$.

- ☐ For every (i,j), $\Sigma_{i,j} = 0$
- ☐ For every (i,j), $\Sigma_{i,j} \geq 0$
- ☑ For every (i,j), $\Sigma_{i,j} = 0 if i \neq j$
- ☑ For all i, $\Sigma_{i,i} \geq 0$

Yes, the answer is correct.
Score: 2

Accepted Answers:
For every (i,j), $\Sigma_{i,j} \geq 0$
For every (i,j), $\Sigma_{i,j} = 0 if i \neq j$
For all i, $\Sigma_{i,i} \geq 0$

8) Given the following ***cumulative distribution*** function, find expected value and variance. ***2 points***

$x \in [1, \infty), F(x) = 1 - \frac{1}{x^3}$

- 6, 3
- 1.5, 0.75
- 1.5, 1.5
- 0.5, 1

Yes, the answer is correct.
Score: 2

Accepted Answers:
1.5, 0.75

9) In this question, we'll take a look at **Conjugate Distributions** and **Conjugate Priors**. ***3 points***
According to Bayesian Probability theory, if the posterior distributions $p(\theta|x)$ are in the same family as the prior probability distribution $p(\theta)$, the prior and posterior are then called **conjugate distributions,** and the prior is called a **conjugate prior** for the likelihood function. As it turns out, the Binomial and Beta Distributions form a pair of Conjugate Distributions.
Pre-requisite Knowledge:

- An *Uninformative Distribution* provides no information at all. In other words, it is an uniform distribution.
- A coin is *fair* if the probability of it turning up heads or tails is exactly 0.5.

Based on your understanding of the above information and the lectures, answer the following question(s).

You are presented with a two-faced coin. To find test if the coin is fair, you decide to toss the coin 10 times. You observe 7 heads and 3 tails in your experiment. Assuming you had no previous knowledge of coins, i.e. assuming an uninformative prior of the appropriate form, what kind of posterior distribution would you have about the coin's fairness? What parameters would this distribution have?

- Binomial Distribution with parameter *p=0.7*
- Normal Distribution with $\mu = 0.7, \sigma = 1.0$
- Beta Distribution with $\alpha = 7, \beta = 3$
- Beta Distribution with $\alpha = 8, \beta = 4$

No, the answer is incorrect.
Score: 0

Accepted Answers:
Beta Distribution with $\alpha = 8, \beta = 4$

10) Prof. Yum Ell has discovered a new method to quickly cool 10ml of a liquid to 4 ^{o} celsius but it appears to be slightly error-prone and thus, the temperature of the final liquid is a random sample from normal distribution $N(4, 1)$. ***3 points***

Prof. Yum Ell now uses his method on 10ml of two liquids A and B (temperatures X and Y respectively) and then mixes these two together. What is the probability that the temperature of the final liquid is more than 4^{o} celsius? What is the variance of the temperature final liquid?

[Assume that the final temperature is the average of the two individual liquids, $Z = \frac{(X+Y)}{2}$]

- 0.5, 1
- 0.5, 0.5
- 0.25, 0.5
- 0.25, 1

Yes, the answer is correct.
Score: 3

Accepted Answers:
0.5, 0.5

Solutions to Quiz #1 can be found here.

Previous Page

End

Unit 6 - Linear Regression

Quiz #2

The due date for submitting this assignment has passed. **Due on 2017-02-08, 23:59 IST.**

Assignment submitted on 2017-02-08, 16:16 IST

This Quiz covers the following Units.

- Linear Algebra
- Statistical Decision Theory
- Linear Regression

This Quiz consists of 10 questions.

- Questions 1-3 carry 1 point each.
- Questions 4-8 carry 2 points each.
- Questions 9-10 carry 3 points each.

The deadline for this Quiz is **08-Feb-2017 11:59PM.**

All the best!

1) Let A be an $m \times n$ matrix of real numbers. The matrix AA^T has an eigenvector $\mathbf{v}$ with eigenvalue λ. Which of the following (eigenvector, eigenvalue) pairs are valid for A^TA ? ***1 point***

- (●) $(A^T\mathbf{v}, \lambda)$
- ( ) $(\mathbf{v}^TA, \lambda)$
- ( ) $(\mathbf{v}, \lambda)$
- ( ) $((A^TA)^{-1}A^T\mathbf{v}, \lambda)$

Yes, the answer is correct.
Score: 1

Accepted Answers:
$(A^T\mathbf{v}, \lambda)$

2) The following is 2D distribution of points and their respective labels (4 classes). ***1 point***

To use the method of regression for classification, we need to first represent the respective labels as a binary vector. Each component of the vector is a separate target.

Assuming the regression is linear, which of the following target mapping schemes can practically fit the data?

- [ ] Standard Binary Code:

0 -> 0,0
1 -> 0,1
2 -> 1,0
3 -> 1,1

- [ ] Gray Code:

0 -> 0,0
1 -> 0,1
2 -> 1,1
3 -> 1,0

- [x] One-hot Code:

0 -> 0,0,0,1
1 -> 0,0,1,0
> 0,1,0,0

3 -> 1,0,0,0

Partially Correct.
Score: 0.5

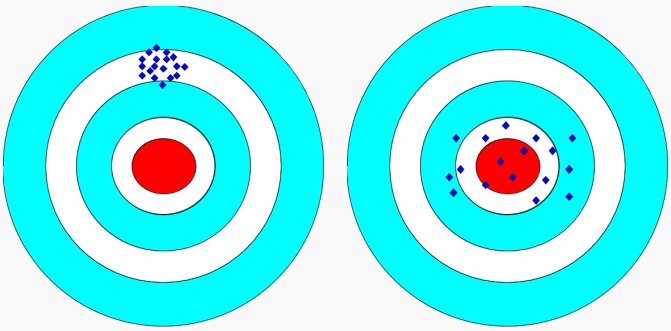
Accepted Answers:
Standard Binary Code:

0 -> 0,0
1 -> 0,1
2 -> 1,0
3 -> 1,1

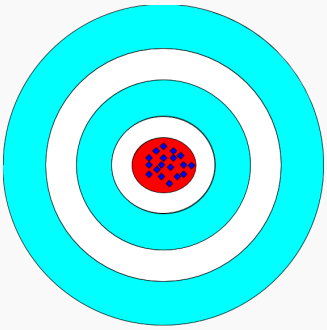
One-hot Code:

0 -> 0,0,0,1
1 -> 0,0,1,0
2 -> 0,1,0,0
3 -> 1,0,0,0

3) Bias and Variance can be visualized using a classic example of a dart game. We can think of the true value of the parameters as the bull's-eye on a target, and the arrow's value as the estimated value from each sample. ***1 point***
Consider the following situations, and select the correct option(s)



Board of player 1 Board of player 2



Board of player 3

- [x] Player 1 has low variance compared to player 2
- [ ] Player 1 has higher variance compared to player 2
- [ ] Bias exhibited by player 2 is clearly more than that done by player 3.

Yes, the answer is correct.
Score: 1

Accepted Answers:
Player 1 has low variance compared to player 2

4) From the lectures, we observe that the Expected Prediction Error, EPE for a model $\hat{f}$, at any point $\mathbf{X}_0$ ***2 points***
can be written as $\mathrm{E}[(\mathrm{E}[\hat{f}(\mathbf{X}_0)] - \hat{f}(\mathbf{X}_0))^2] + [\mathrm{E}[\hat{f}(\mathbf{X}_0)] - f(\mathbf{X}_0)]^2 + \sigma^2$.
The first term seems to have an expectation of some sorts. What distribution is it an expectation over and why?

- ○ Both the inner and the outer expectations are over the distribution from which the dataset is drawn from. Variance in the dataset will contribute to the EPE.
- ○ The inner expectation is over the dataset and the outer expectation is over the model space(i.e. distribution of all models). Difference between how the data varies inherently and how the model learns the distribution

should contribute to the EPE.

- (•) Both the inner and outer expectations are over the model space. The term, as a whole, represents the variance in the estimate when using different models(trained on different datasets drawn from the same distribution). This inherent variance in the model should contribute to the EPE.

Yes, the answer is correct.
Score: 2

Accepted Answers:
Both the inner and outer expectations are over the model space. The term, as a whole, represents the variance in the estimate when using different models(trained on different datasets drawn from the same distribution). This inherent variance in the model should contribute to the EPE.

5) Consider a modified $k-NN$ method in which once the $k-$ nearest neighbours to the query point are identified, you fit a linear model(i.e. learn a linear regression model over them) and compute the output value for the query point using the fitted model. ***2 points***
Which of the following are correct?

- [ ] This method makes an assumption that the data is locally linear.
- [x] In order to perform well, this method would need densely distributed training data.
- [ ] This method has higher bias compared to $k-NN$.
- [x] This method has higher variance compared to $k-NN$.

No, the answer is incorrect.
Score: 0

Accepted Answers:
This method makes an assumption that the data is locally linear.
In order to perform well, this method would need densely distributed training data.
This method has higher bias compared to $k-NN$.

6) Students of XYZ University are attending an ML contest. They presented with some data $\mathbf{X}$. Without analyzing the dataset, **Team Ov3rConfyd3nt** decides to try to fit a linear model on the dataset. What results await **Team Ov3rConfyd3nt** if $\mathbf{X}$ is not full-rank? ***2 points***

- ( ) Full-rank or not, there exists a simple formula that can give the linear regression coefficients in one step. If they find high error, they'll now have more time to try out different models.
- (•) Since the data matrix is not full-rank, their code will throw them an error because $\mathbf{X}^T\mathbf{X}$ is singular. You cannot invert a singular matrix.
- ( ) Since the data matrix is not full-rank, their feature vectors are not linearly independent. This will give them wrong coefficients.

Yes, the answer is correct.
Score: 2

Accepted Answers:
Since the data matrix is not full-rank, their code will throw them an error because $\boldsymbol{X}^T\boldsymbol{X}$ is singular. You cannot invert a singular matrix.

7) In this question, we explore ***asymmetric loss functions***. In certain situations, erroneously classifying an element of **class A** as **class B** is more harmful than classifying **class B** as **class A.** ***2 points***
Consider the case of a chip-making factory which manufactures chipsets for mission-critical applications. From a large number of chips, the testing mechanism must check each chip for flaws and classify each chip as good/bad.
Which of the following ***Loss Matrices*** is most appropriate for this task?
Note: The matrix entry ***L(i,j)*** refers to the cost if an entity of class **i** is predicted to be class **j** and class 0 refers to **bad chip** while class 1 is for **good chips**.

- ( )
$$\begin{pmatrix} 0.1 & 1 \\ 1 & 0 \end{pmatrix}$$
- (•)
$$\begin{pmatrix} 0 & 1 \\ 0.2 & 0 \end{pmatrix}$$
- ( )
$$\begin{pmatrix} 0 & 0.5 \\ 1 & 0 \end{pmatrix}$$
- ( )
$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

Yes, the answer is correct.
Score: 2

Accepted Answers:

$$\begin{pmatrix} 0 & 1 \\ 0.2 & 0 \end{pmatrix}$$

8) Which of the following statements about ***eigenvectors*** are true? ***2 points***

- [ ] 3x3 Orthogonal matrices (that are *not* Identity) do not have non-trivial eigenvectors
- [ ] 2x2 Orthogonal matrices (that are *not* Identity) do not have non-trivial eigenvectors
- [x] Let **x** and **y** be eigenvectors of **A** such that both have the *same* eigenvalue. a**x**+b**y** is also an eigenvector.
- [ ] Let **x** and **y** be eigenvectors of **A.** For *a, b > 0* and *a+b = 1*, a**x**+b**y** is also an eigenvector.

Yes, the answer is correct.
Score: 2

Accepted Answers:
*Let **x** and **y** be eigenvectors of **A** such that both have the same eigenvalue. a**x**+b**y** is also an eigenvector.*

9) In the lecture on Multivariate Regression, you learn about using orthogonalization iteratively to obtain regression co-effecients. This method is generally referred to as Multiple Regression using Successive Orthogonalization. ***3 points***

In the formulation of the method, we observe that in iteration *k*, we regress the entire dataset on $\mathbf{z}_0, \mathbf{z}_1, \ldots \mathbf{z}_{k-1},$. It seems like a waste of computation to recompute the coefficients for $\mathbf{z}_0$ a total of *p* times, $\mathbf{z}_1$ a total of *p-1* times and so on. Can we re-use the coefficients computed in iteration *j* for iteration *j+1* for $\mathbf{z}_{j-1}$?

- ( ) No. Doing so will result in the wrong γ matrix. and hence, the wrong β_i's.
- (•) Yes. Since $\mathbf{z}_{j-1}$ is orthogonal to $\mathbf{z}_{j-l} \forall l \leq j-1$, the multiple regression in each iteration is essentially a univariate regression on each of the previous residuals. Since the regression coefficients for the previous residuals don't change over iterations, we can re-use the coefficients for further iterations.

Yes, the answer is correct.
Score: 3

Accepted Answers:
Yes. Since $\mathbf{z}_{j-1}$ is orthogonal to $\mathbf{z}_{j-l} \forall l \leq j-1$, the multiple regression in each iteration is essentially a univariate regression on each of the previous residuals. Since the regression coefficients for the previous residuals don't change over iterations, we can re-use the coefficients for further iterations.

10) Let **A** of size *nxn* and **B** of size *n*xn be two matrices such that rank(**A**) < a and rank(**B**) < b. ***3 points***

- [ ] rank(**AB**) <= min(a,b)
- [x] $\mathbf{A}^{-1}$ is not defined
- [x] $\mathbf{A}^T\mathbf{A} \neq \mathbf{I}$
- [ ] **A** has multiple vanishing (zero) singular values.

Partially Correct.
Score: 1.5

Accepted Answers:
*rank(**AB**) <= min(a,b)*
$\mathbf{A}^{-1}$ is not defined
$\mathbf{A}^T\mathbf{A} \neq \mathbf{I}$
***A** has multiple vanishing (zero) singular values.*

Solutions to Quiz #2 can be found here.

Unit 8 - Classification - Linear Models

Programming Assignment #1

The due date for submitting this assignment has passed.
As per our records you have not submitted this assignment.

Due on 2017-02-22, 23:59 IST.

Welcome to **Programming Assignment #1**
The purpose of this assignment is to get you acquainted with a very popular Machine Learning tool, **Weka**.

Before attempting this assignment, complete the following checklist.

- Watch the tutorial lecture on Weka. This is located under **Classification - Linear Models**.
- Download Weka from http://www.cs.waikato.ac.nz/ml/weka/downloading.html.
- Watch all the lectures scheduled for Week #3. This includes **Dimensionality Reduction** and **Classification-Linear Models**.
- Complete **Quiz #3** to test your understanding of topics.
- Download the datasets required for this assignment from here.

You are now ready to begin this programming assignment. Good luck!

Dataset #1
Files - dataset1.csv, dataset1.arff, dataset1_test.arff, dataset1_train.arff
This is a synthetic dataset to get you started with Weka. If you look into the arff/csv files, you'll notice that the data consists of 1000 (X,Y) pairs with 3-dimensional X and 1-dimensional Y.
To get you started, each of the datasets has already been split into a train and test dataset. **Please note that your test dataset should not be used for anything other than computing the error of your model. Use only the train dataset to adjust parameters.**
To load a custom dataset for testing, under Classify tab, select 'Supplied test set -> Open File' to load your test dataset. All evaluations Weka performs are on the test dataset.

Tasks
You are to train a Linear Regression Model on the data. Play around with different parameters for the model.
Note: The following parameters are to be fixed.

- Ridge coefficient = 0
- Attribute Selection = None
- Collinear Attribute Elimination = False

Now answer the following questions. You may need to use the tool and do some extra work in the process.

1) What is the best *unregularized* linear fit obtained for this dataset? ***1 point***
Select the option closest to the $(\beta_0, \beta_1, \beta_2, \beta_3)$ you obtained through Weka.

- (4, 3, 4, 5)
- (2, 3, 4, 5)
- (2.7, 4.4, 3.56, 5.75)
- (0.02, 1.98, 6.74, -1.43)

No, the answer is incorrect.
Score: 0

Accepted Answers:
(4, 3, 4, 5)

2) On the test data, what values do you obtain for the best model for $(N_{train}, N_{test}, RMS_{test})$? ***1 point***
Note : RMS stands for Root Mean Squared (Error) and N_X is the number of data points in the dataset ***X***.

- (599, 501, 0.002)
- (599, 401, 0.0002)
- (401, 599, 0.0003)
- (599, 401, 0.0003)

No, the answer is incorrect.
Score: 0

Accepted Answers:
(599, 401, 0.0003)

3) Upon careful inspection of the errors, we find that they are relatively small and the data has been synthetically generated with very low noise. In such a scenario, what would happen if we employed Ridge Regression with $\lambda = 100$ instead of Linear Regression (same as Ridge Regression with $\lambda = 0$)? ***1 point***

- ☐ Ridge Regression will shrink coefficients to 0.
- ☐ Because the error is already close to 0, Ridge Regression will not perform any shrinkage.
- ☐ The errors will be higher than currently obtained.
- ☐ The errors will be lower than currently obtained.

No, the answer is incorrect.
Score: 0

Accepted Answers:
Ridge Regression will shrink coefficients to 0.
The errors will be higher than currently obtained.

Dataset #2 - AutoMPG

Files - auto* (* is any string)
This dataset is the AutoMPG dataset from UCI repository. You are provided with the data in ARFF format. Upon loading the dataset, you'll observe that it contains attributes from both continuous and discrete domains.

Tasks

As with the other datasets, you are to train a Linear Regression model on the given data. Before you do that, you'll first convert the discrete-valued attributes to binary form.

- One Hot Encoding - Under the Preprocess tab, select Filter -> Choose -> weka/filters/unsupervised/attribute/NominalToBinary, you have now selected the **Nominal to Binary converter**. Upon applying the filter, you should now notice an increase in the number of attributes. Note the number of attributes before and after.
- Regressor - Train a Linear Regression model with the parameter settings as mentioned in Dataset #1. You will have to vary the Ridge coefficient for the questions that follow.
- Derive insights into the data looking at graphs generated by **Preprocess->Visualize All** and **Visualize**.

Now answer the following questions. You may need to use the tool and do some extra work in the process.

4) What are the number of attributes before and after applying *NominalToBinary* on the given data? ***1 point***

- ○ (7, 26)
- ○ (8, 26)
- ○ (8, 22)
- ○ (7, 22)

No, the answer is incorrect.
Score: 0

Accepted Answers:
(8, 26)

5) What happens to the constant in the regression coefficients set as you increase λ(in Ridge Regression) from 0 to 10000? ***1 point***

- ○ It converges to $\beta_0 = \bar{y} = \frac{\sum_{i=1}^{N} y_i}{N}$ always.
- ○ It converges to $\beta = \sum_{i=1}^{N} y_i$ always.
- ○ It might never converge to a value.
- ○ As observed empirically, the value oscillates between negative and positive values. It will always converge to 0 as this is the solution obtained by the optimization problem solved by Ridge Regression.

No, the answer is incorrect.
Score: 0

Accepted Answers:
It converges to $\beta_0 = \bar{y} = \frac{\sum_{i=1}^{N} y_i}{N}$ always.

6) Now execute *CVParameterSearch* on the data with the above given configuration. Has your model become better or worse? What is the λ returned by *CVParameterSearch* ? ***1 point***

- ○ Better with $\lambda \approx 2.85$
- ○ Worse with $\lambda \approx 2.97$
- ○ Better with $\lambda \approx 1.5$

- Worse with $\lambda \approx 1.5$

No, the answer is incorrect.
Score: 0

Accepted Answers:
Worse with
$\lambda \approx 2.97$

Dataset #3 - Prostate Dataset
Files - prostate* (* is any string)
This is the Prostate Cancer dataset talked about in *Elements of Statistical Learning.* Read the info file for more information on the dataset. Use the train dataset for training and the test dataset to report numbers on the model.

Tasks
As with the previous dataset, you'll train a Linear Regression model on the given data. You will also perform *hyperparameter* search using Weka built-ins and report numbers on the same.

- Learn a Linear Regression Model with the same settings as the previous question.
- Now perform a hyperparameter search for the Ridge Coefficient by using *CVParameterSearch* in the meta functions (instead of *LinearRegression*). Feed in *LinearRegression* as the model you're trying to learn. Find a good Ridge Coefficient (R) for the LinearRegression Model over the range 0 - 50 with 100 different values and 5-fold cross-validation. Note down your best R and the corresponding error.

Now answer the following questions. You may need to use the tool and do some extra work in the process.

7) When you run *Visualize All* (Preprocess -> Visualize All) on the data, what attribute value distributions appear to be similar to that of the target attribute? Based on only this knowledge and nothing else, can you consider only these attributes for regression and rule out others as they have visually different distributions? ***1 point***

- ☐ **lcavol, lweight**
- ☐ **lbph, svi**
- ☐ **age**
- ☐ Since only these attributes have similar distributions, we can throw away all the others safely.
- ☐ No. We cannot perform attribute selection in such a manner. Transformation from the source attribute to the target attribute could be non-linear transformation. Other attributes might be able to correlate changes in target attribute with non-linear transformations as well.

No, the answer is incorrect.
Score: 0

Accepted Answers:
lcavol, lweight
age
No. We cannot perform attribute selection in such a manner. Transformation from the source attribute to the target attribute could be non-linear transformation. Other attributes might be able to correlate changes in target attribute with non-linear transformations as well.

8) Looking at the plots under the **Visualize** tab, pay close attention to ***X*** *vs* ***svi*** and ***svi*** *vs* ***Y*** plots (X and Y are other attributes). What can you conclude about the nature of **svi** ? ***1 point***

- ☐ **svi** is a binary-valued attribute.
- ☐ **svi** has values only in the range **[0,1]**.
- ☐ **svi** is highly correlated with the output.
- ☐ **svi** is a continuous-valued attribute.

No, the answer is incorrect.
Score: 0

Accepted Answers:
svi *is a binary-valued attribute.*
svi *has values only in the range* ***[0,1]****.*

9) How does your model far when the Ridge Coefficient is changed 20 from 0? ***1 point***

- The model does better in a few metrics, but also gives higher errors.
- The model does better in all the metrics shown on Weka.
- No noticeable change is seen.

No, the answer is incorrect.
Score: 0

Accepted Answers:
The model does better in all the metrics shown on Weka.

10) If you plot Cross Validation Error (y) vs Ridge Coefficient (x), what kind of a curve would you expect? ***1 point***

- Straight Line passing though the origin.
- U-shaped curve with a minimum somewhere in between.
- Inverted-U shaped curve with a maximum in between.
- Horizontal Line i.e. No change in CV Error with change in Ridge Coefficient

No, the answer is incorrect.
Score: 0

Accepted Answers:
U-shaped curve with a minimum somewhere in between.

Previous Page

Next Page

Unit 8 - Classification - Linear Models

Quiz #3

The due date for submitting this assignment has passed.
As per our records you have not submitted this assignment.

Due on 2017-02-15, 23:59 IST.

This quiz will test the following Units.

- Dimensionality Reduction
  - Subset Selection 1 & 2
  - Shrinkage Methods
  - Principal Component Regression
  - Partial Least Squares
- Classification - Linear Models
  - Linear Classification
  - Logistic Regression
  - Linear Discriminant Analysis 1, 2 & 3

This quiz consists of 11 questions with the marking scheme indicated against each question.

Good Luck!

1) Between Subset Selection and Shrinkage Methods, what class of methods will you prefer to use on large datasets for dimensionality reduction, given that you have no prior domain knowledge? ***1 point***

- ○ Subset Selection Methods.
- ○ Shrinkage Methods
- ○ Either; It doesn't matter as both get the job done.

No, the answer is incorrect.
Score: 0

Accepted Answers:
Shrinkage Methods

2) Which of the following dimensionality reduction methods are *hard thresholding* methods? ***1 point***
(A method is soft-thresholding if it contains parameters that vary in a continuous fashion.)

- ☐ Forward Stepwise Regression
- ☐ Backward Stepwise Regression
- ☐ Forward Stagewise Regression
- ☐ Ridge Regression
- ☐ Least Absolute Shrinkage and Selection Operator (LASSO)
- ☐ Principal Component Regression
- ☐ Partial Least Squares Method

No, the answer is incorrect.
Score: 0

Accepted Answers:
Forward Stepwise Regression
Backward Stepwise Regression
Forward Stagewise Regression
Principal Component Regression
Partial Least Squares Method

3) What happens to the solution returned by LASSO when $t \geq \sum_{m=1}^{p} |\beta_m^{LS}|$? ***1 point***
Note :

- $\forall m, \beta_m^{LS}$ is the coefficient of $\mathbf{X}_i$ in the solution returned by standard Least Squares Method.
- RSS - Residual Sum of Squares

○
LASSO solution converges to the Least Squares solution with $RSS_{LASSO} = RSS_{LS}$ after LASSO is run.
○

LASSO returns a different solution with $RSS_{LASSO} < RSS_{LS}$ after LASSO is run.

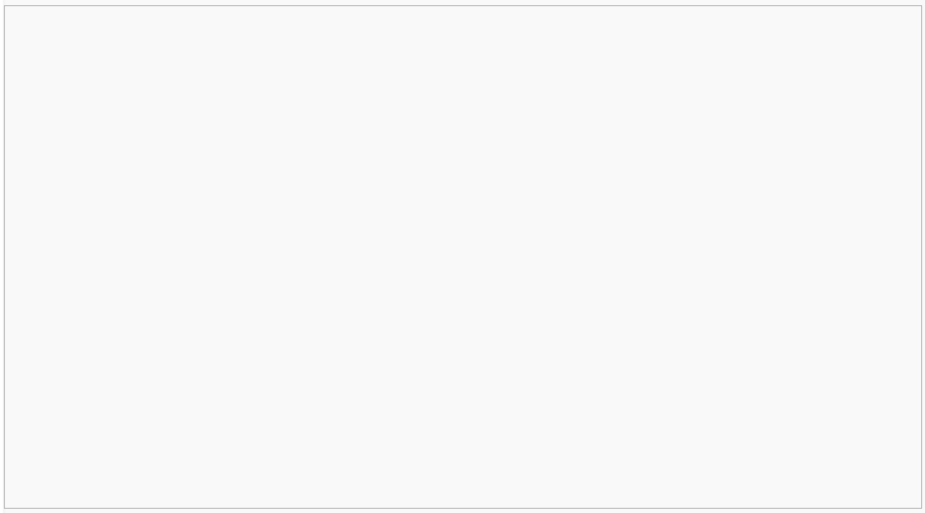
LASSO returns a different solution with $RSS_{LASSO} > RSS_{LS}$ after LASSO is run.

Since LASSO solves an optimization problem, it shrinks some weights to Zero regardless of the value assigned to t or λ. Nothing can be said about the RSS of LASSO relative to RSS of LS.

No, the answer is incorrect.
Score: 0

Accepted Answers:
LASSO solution converges to the Least Squares solution with $RSS_{LASSO} = RSS_{LS}$ after LASSO is run.

4) In the following dataset, there are two classes arranged in the following manner: ***1 point***



Which of the following bases functions would you use to prevent any masking?
Select all that apply.

- $1, x, x^3, x^4$
- $1, x, sin(x)$
- $1, x, x^2$
- $1, x, x^2, x^3$

No, the answer is incorrect.
Score: 0

Accepted Answers:
$1, x, x^3, x^4$
$1, x, sin(x)$

5) Given the following distribution of data points: ***1 point***

What method would you choose to perform *Dimensionality Reduction?*

- ○ Linear Discriminant Analysis.
- ○ Principal Component Analysis

No, the answer is incorrect.
Score: 0

Accepted Answers:
Linear Discriminant Analysis.

6) Select the valid reasons for doing dimensionality reduction. ***2 points***

- ☐ Variance can be controlled by controlling the number of variables used in the model.
- ☐ Lower number of variables render the model to better interpretability.
- ☐ Upon reducing the variable pool, we can sometimes increase the prediction accuracy, though not always.
- ☐ Upon reducing the variable pool, we can reduce the time taken to run inference on the model at the cost of increased model building time.

No, the answer is incorrect.
Score: 0

Accepted Answers:
Variance can be controlled by controlling the number of variables used in the model.
Lower number of variables render the model to better interpretability.
Upon reducing the variable pool, we can sometimes increase the prediction accuracy, though not always.
Upon reducing the variable pool, we can reduce the time taken to run inference on the model at the cost of increased model building time.

7) Regular **LDA** does not make a Naive-Bayes assumption; i.e. the Gaussian model assumed need not be a product of independent Gaussians. Assume that we have a new classifier called the **NB** classifier which is identical to **LDA** but also contains the Naive-Bayes assumption. **NB** Gaussian model can be expressed as a product of univariate Gaussians of each individual component. ***2 points***
Which of the following statements are ***true***?

- ☐ If the covariance matrices of all classes are the same, the separating surface is a hyperplane (linear).
- ☐ The covariance matrix of every class is necessarily ***orthogonal***
- ☐ The covariance matrix of every class is necessarily ***diagonal***
- ☐ The Gaussian Model expressed by **NB** is necessarily a **spherical Gaussian**.

No, the answer is incorrect.
Score: 0

Accepted Answers:
If the covariance matrices of all classes are the same, the separating surface is a hyperplane (linear).
The covariance matrix of every class is necessarily ***diagonal***

8) Which of the following statements are **true** about **Linear Discriminant Analysis?** ***2 points***

- ☐

If the covariance matrix is $k.I$, $w \propto (m_1 - m_2)$

If the covariance matrix is **diagonal,** $w \propto (m_1 - m_2)$

If $w \propto (m_1 - m_2)$, the covariance matrix is **diagonal.**

If $w \propto (m_1 - m_2)$, the covariance matrix is $k.I$

No, the answer is incorrect.
Score: 0

Accepted Answers:

If the covariance matrix is $k.I$, $w \propto (m_1 - m_2)$

9) In general, which of the following classification methods is the most resistant to ***gross outliers***? ***2 points***

- Linear Discriminant Analysis (LDA)
- Quadratic Discriminant Analysis (QDA)
- Logistic regression
- Linear Regression

No, the answer is incorrect.
Score: 0

Accepted Answers:
Logistic regression

10) What is the effect of increasing λ on coefficients of the basis vectors of $\mathbf{X}$ in Ridge Regression ? ***3 points***

- It shrinks the Principal Components of $\mathbf{X}$ with higher variance **more** than those with lower variance.
- It shrinks the Principal Components of $\mathbf{X}$ with lower variance **more** than those with higher variance.
- It shrinks the $\mathbf{X}_i$'s with lower variance **more** than those with higher variance.
- It shrinks the $\mathbf{X}_i$'s with higher variance **more** than those with lower variance.
- It has no effect on any of the basis vectors as the optimization problem it is solving isn't affected by any statistical property of any of the basis vectors.

No, the answer is incorrect.
Score: 0

Accepted Answers:
*It shrinks the Principal Components of X with lower variance **more** than those with higher variance.*

11) Consider data-points with a single dimension and belonging to one of two possible classes. ***3 points***
The data points are arranged on a line in such a way that they are fully separated by a point x_0, such that points belonging to class 1 & 2 lie entirely on opposite sides of the point.
Now, if we use Logistic Regression to fit a line $y = \beta_0 + \beta.x$, what are β and β_0

- $1, x_0$
- ∞, x_0
- ∞, ∞
- $0, x_0$

No, the answer is incorrect.
Score: 0

Accepted Answers:
∞, x_0

Unit 10 - Classification - Separating Hyperplane Approaches

Quiz #4

The due date for submitting this assignment has passed.
As per our records you have not submitted this assignment.

Due on 2017-02-23, 23:55 IST.

1) Given the following distribution of 2 classes: *1 point*

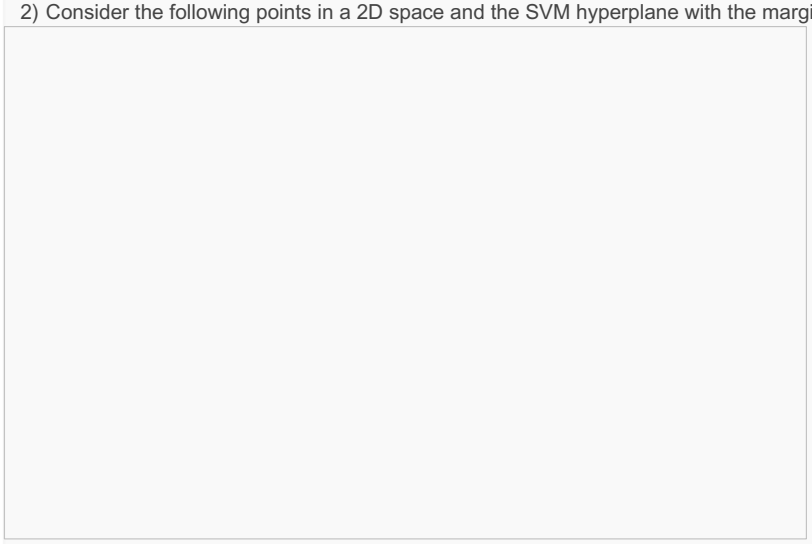
Which of the following linear classification methods will produce the **worst** boundary?

- SVM-C (non-zero C)
- Logistic Regression
- SVM, C=0
- Linear Discriminant Analysis

No, the answer is incorrect.
Score: 0

Accepted Answers:
SVM, C=0

2) Consider the following points in a 2D space and the SVM hyperplane with the margins: *2 points*



Which of the following points can be removed without affecting the hyperplane?

- [ ] I
- [ ] III
- [ ] II
- [ ] IV

No, the answer is incorrect.

Score: 0

Accepted Answers:
III

3) Consider the situation where the **hinge-loss** function at the end of the **SVM: Hinge Loss Formulation** ***3 points***
lecture was replaced with the **logistic loss** function (also called 'softplus').
Which of the following statements would be **TRUE**?

- ☐ α's are not 0 for points beyond the margin on the correct side of the hyper-plane.
- ☐ α's are 0 for points beyond the margin on the correct side of the hyper-plane.
- ☐ For points that are on the correct side of the hyperplane, but within the margin($0 < y_i f(x_i) < 1$), the loss contributed by a point that is **twice** as far away from the **margin** (relative to some other point) is **more than** 2 times it's loss.
- ☐ The resulting optimisation problem is not convex.

No, the answer is incorrect.
Score: 0

Accepted Answers:
α's are not 0 for points beyond the margin on the correct side of the hyper-plane.
*For points that are on the correct side of the hyperplane, but within the margin($0 < y_i f(x_i) < 1$), the loss contributed by a point that is **twice** as far away from the **margin** (relative to some other point) is **more than** 2 times it's loss.*

4) Which of the following statements about Kernel Functions are **TRUE?** Assume, in each case, that the ***2 points***
vector **x** has 2 dimensions.

- ☐ The implicit vector transformation for the kernel $K(x,x') = (1 + < x,x' >)^4$ has 9 dimensions.
- ☐ The implicit vector transformation for the kernel $K(x,x') = (< x,x' >)^4$ has 3 dimensions.
- ☐ The implicit vector transformation for the kernel $K(x,x') = tanh(< x,x' >)$ has ∞ dimensions.

No, the answer is incorrect.
Score: 0

Accepted Answers:
The implicit vector transformation for the kernel $K(x,x') = (< x,x' >)^4$ has 3 dimensions.
The implicit vector transformation for the kernel $K(x,x') = tanh(< x,x' >)$ has ∞ dimensions.

5) ***1 point***

$$y = a_4x^4 + a_3x^3 + a_2x^2 + a_1x^1 + a_0$$

Give the above polynomial, check whether the function is convex for each of the following sets of coefficients:

- ☐ a4=1, a3=4, a2=6, a1=4, a0=1
- ☐ a4=1, a3=-2, a2=1, a1=0, a0=0
- ☐ a4=0, a3=0, a2=1, a1=2, a0=1
- ☐ a4=0, a3=1, a2=3, a1=2, a0=-1

No, the answer is incorrect.
Score: 0

Accepted Answers:
a4=1, a3=4, a2=6, a1=4, a0=1
a4=0, a3=0, a2=1, a1=2, a0=1

6) Which of the following optimisation problems does **not** satisfy Slater's condition? ***1 point***

- ☐ $min(x^2 + y^2)$
  subject to
  $(x-1)^2 + y^2 \leq 1$
  and
  $(x+1)^2 + y^2 \leq 1$
- ☐ $minx(x-1)(x-2)$
  subject to
  $x^2 + y^2 \leq 6$
- ☐ $min(x^2)$
  subject to
  $x^2 + y^2 \leq 4$

and

$y = 1$

☐

$min((x-4)^2 + y^2)$

subject to

$x^2 + \frac{y^2}{4} \leq 4$

and

$x \leq 1$

No, the answer is incorrect.
Score: 0

Accepted Answers:

$min(x^2 + y^2)$
subject to
$(x-1)^2 + y^2 \leq 1$
and
$(x+1)^2 + y^2 \leq 1$
$minx(x-1)(x-2)$
subject to
$x^2 + y^2 \leq 6$

7) Answer the questions that follow with respect to the following convex optimisation problem. $min(x^2 + y^2)$ ***2 points***
subject to
$(x-2)^2 + y^2 < 1$

☐ The duality gap is 0.

☐

The primal solution is $(x, y) = (1, 0)$

☐

The dual problem is $sup_\lambda(\frac{4\lambda^2}{(1+\lambda)^2} + \lambda(\frac{4}{(1+\lambda)^2} - 1))$

☐

The dual solution is $(x, y, \lambda) = (1, 0, -1)$

No, the answer is incorrect.
Score: 0

Accepted Answers:

The duality gap is 0.
The primal solution is $(x, y) = (1, 0)$
The dual problem is $sup_\lambda(\frac{4\lambda^2}{(1+\lambda)^2} + \lambda(\frac{4}{(1+\lambda)^2} - 1))$

Previous Page

End

Unit 12 - Parameter Estimation

Quiz #5

The due date for submitting this assignment has passed. **Due on 2017-03-05, 23:59 IST.**

Assignment submitted on 2017-03-02, 00:01 IST

Welcome to Quiz #5.
This Quiz covers the following units.

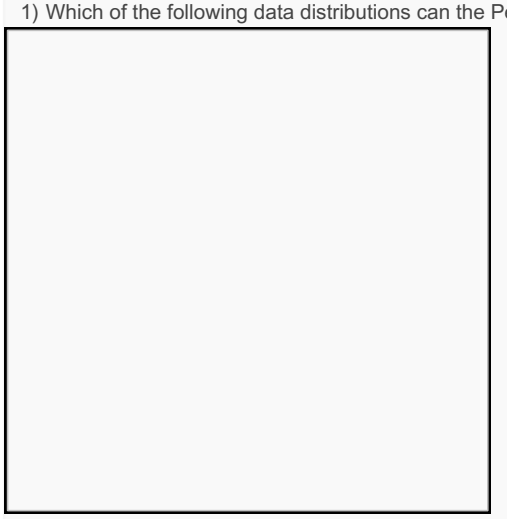
1. Artificial Neural Networks (ANNs)
2. Parameter Estimation

Read the following before you begin.

- This Quiz consists of 10 questions
- Questions 1-4 carry 1 points each; 5-8 carry 2 points each and 9-10 carry 3 points each
- You may require a calculator for this quiz
- This Quiz is open for 7 days and will stop accepting submissions at **11:59PM on 01-Mar-2017.**

Good Luck!

1) Which of the following data distributions can the Perceptron learn to classify with **zero error**? ***1 point***



- [ ] (b)
- [x] (d)
- [x] (a)
- [x] (c)

Yes, the answer is correct.
Score: 1

Accepted Answers:
(d)
(a)
(c)

2) Which of the following functions can be used as an activation function in Artificial Neural Networks? ***1 point***

- [x] $\sigma(x) = \frac{1}{1+e^{-x^3}}$
- [ ] $\sigma(x) = arctan(\frac{x^4}{2})$
- [x] $\sigma(x) = 1 - sgn(x+3)$
- [ ]

$\sigma(x) = 1 - |x^2|$

No, the answer is incorrect.
Score: 0

Accepted Answers:

$\sigma(x) = \frac{1}{1+e^{-x^3}}$

$\sigma(x) = arctan(\frac{x^4}{2})$

$\sigma(x) = 1 - |x^2|$

3) Which of the following are true if your prior is very different from the actual distribution when estimating parameters? ***1 point***

- [x] We might end up with wrong parameters.
- [x] Any errors can be corrected for by using more data.
- [x] Any errors can be corrected for by adjusting the prior to resemble the actual distribution more.
- [ ] Any errors can be corrected for by using a smaller (random) subset of data instead of using full data for paramter estimation.

Yes, the answer is correct.
Score: 1

Accepted Answers:
We might end up with wrong parameters.
Any errors can be corrected for by using more data.
Any errors can be corrected for by adjusting the prior to resemble the actual distribution more.

4) Given N samples $x_1, x_2, \ldots, x_N$ drawn independently from a Gaussian distribution with variance σ^2 and unknown mean μ, find the MLE of the mean. ***1 point***

- ○ $\mu_{MLE} = \frac{\sum_{i=1}^{N} x_i}{\sigma^2}$
- ◉ $\mu_{MLE} = \frac{\sum_{i=1}^{N} x_i}{2\sigma^2 N}$
- ○ $\mu_{MLE} = \frac{\sum_{i=1}^{N} x_i}{N}$
- ○ $\mu_{MLE} = \frac{\sum_{i=1}^{N} x_i}{N-1}$

Yes, the answer is correct.
Score: 1

Accepted Answers:

$\mu_{MLE} = \frac{\sum_{i=1}^{N} x_i}{N}$

5) Continuing with the above question, assume that the prior distribution of the mean is also a Gaussian distribution, but with parameters mean μ_p and variance σ_p^2. Find the MAP estimate of the mean. ***2 points***

- ◉ $\mu_{MAP} = \frac{\sigma^2 \mu_p + \sigma_p^2 \sum_{i=1}^{N} x_i}{\sigma^2 + N\sigma_p^2}$
- ○ $\mu_{MAP} = \frac{\sigma^2 + \sigma_p^2 \sum_{i=1}^{N} x_i}{\sigma^2 + \sigma_p^2}$
- ○ $\mu_{MAP} = \frac{\sigma^2 + \sigma_p^2 \sum_{i=1}^{N} x_i}{\sigma^2 + N\sigma_p^2}$

$$\mu_{MAP} = \frac{\sigma^2\mu_p + \sigma_p^2\sum_{i=1}^{N} x_i}{N(\sigma^2+\sigma_p^2)}$$

Yes, the answer is correct.
Score: 2

Accepted Answers:

$$\mu_{MAP} = \frac{\sigma^2\mu_p + \sigma_p^2\sum_{i=1}^{N} x_i}{\sigma^2+N\sigma_p^2}$$

6) Why do we require the output of every neuron to be in the linear region of its activation function at the start of training? ***2 points***

- [x] Enforcing this condition gives us larger gradients.
- [ ] Enforcing this condition gives us smaller gradients.
- [ ] We require lower variance in outputs in the initial phase of training.
- [x] We require faster learning in the initial phase of learning.

Yes, the answer is correct.
Score: 2

Accepted Answers:
Enforcing this condition gives us larger gradients.
We require faster learning in the initial phase of learning.

7) Which of the following are true when comparing ANNs and SVMs? ***2 points***

- [x] ANN error surface has multiple local minima while SVM error surface has only one minima
- [x] After training, an ANN might land on a different minimum each time, when initialized with random weights during each run.
- [x] In training, ANN's error surface is navigated using a gradient descent technique while SVM's error surface is navigated using convex optimization solvers.
- [ ] As shown for Perceptron, there are some classes of functions that cannot be learnt by an ANN. An SVM can learn a hyperplane for any kind of distribution.

Yes, the answer is correct.
Score: 2

Accepted Answers:
ANN error surface has multiple local minima while SVM error surface has only one minima
After training, an ANN might land on a different minimum each time, when initialized with random weights during each run.
In training, ANN's error surface is navigated using a gradient descent technique while SVM's error surface is navigated using convex optimization solvers.

8) What happens to the error surface when we increase the temperature parameter in Simulated Annealing starting with a very low temperature? ***2 points***
Note : A minimum X is **deeper** than minimum Y if $F(\cdot)|_X \leq F(\cdot)|_Y$.

- ( ) More minima are formed with **deepest** minima first.
- ( ) More minima are formed with **deepest** minima last.
- ( ) More minima disappear with **deepest** first.
- (•) More minima disappear with **shallowest** first.

Yes, the answer is correct.
Score: 2

Accepted Answers:
*More minima disappear with **shallowest** first.*

9) Consider a 3-layer Neural Network with 3 Input features, 3 Hidden Neurons and 1 output neuron (3-3-1 architecture). Evaluate the network output using the following weights and biases. Use $\sigma(x) = \frac{1}{1+e^x}$ and $\sigma(x) = cos^{-1}(x)$ as transfer functions for Layers 2 and 3 respectively. ***3 points***

Weights between Layer 1 and 2 - $\begin{pmatrix} 0.3 & 1.7 & -2.1 \\ 3.1 & -1.6 & 0.2 \\ 0.24 & -0.31 & 0.1 \end{pmatrix}$

Layer 2 Biases - $\begin{pmatrix} 0.1 \\ -0.2 \\ 0.3 \end{pmatrix}$

Weights between Layer 2 and 3 - $\begin{pmatrix} -0.9 \\ -0.7 \\ 0.6 \end{pmatrix}$

Layer 3 Bias - **0.7**

Input - $\begin{pmatrix} 0.7 \\ -0.6 \\ 0.65 \end{pmatrix}$

- 0.656
- 1.5051
- 0.6056
- 1.0551

Yes, the answer is correct.
Score: 3

Accepted Answers:
1.5051

10) For the standard 3-layer neural network architecture discussed in lectures, select the correct expression for δ_{ki} used in evaluation of $\frac{\partial R_i}{\beta_{km}}$ if $g_{ki}(T_i) = \frac{sin(T_{ki})}{\sum_{l=1}^{K} sin(T_{li})}$. ***3 points***

- [x] $\delta_{ki} = -2 \cdot (y_{ik} - f_k(x_i)) \cdot z_{mi} \cdot cos(T_{ki}) \cdot \frac{\sum_{l \neq k} sin(T_{li})}{(\sum_l sin(T_{li}))^2}$
- [ ] $\delta_{ki} = -2 \cdot (y_{ik} - f_k(x_i)) \cdot z_{mi} \cdot g_{ki}(T_i) \cdot (1 - g_{ki}(T_i))$
- [ ] $\delta_{ki} = -2 \cdot (y_{ik} - f_k(x_i)) \cdot z_{mi} \cdot (g_{ki}(T_i))^2 \cdot (1 - (g_{ki}(T_i))^2)$
- [x] $\delta_{ki} = -2 \cdot (y_{ik} - f_k(x_i)) \cdot z_{mi} \cdot \frac{cos(T_{ki})}{\sum_l sin(T_{li})} \cdot \left(1 - \frac{sin(T_{ki})}{\sum_l sin(T_{li})}\right)$

No, the answer is incorrect.
Score: 0

Accepted Answers:

$\delta_{ki} = -2 \cdot (y_{ik} - f_k(x_i)) \cdot z_{mi} \cdot cos(T_{ki}) \cdot \frac{\sum_{l \neq k} sin(T_{li})}{(\sum_l sin(T_{li}))^2}$

Previous Page

End

Unit 14 - Evaluation Measures

Quiz #6

The due date for submitting this assignment has passed. **Due on 2017-03-11, 23:59 IST.**

Assignment submitted on 2017-03-08, 22:05 IST

1) Which of the following statements are true with respect to the application of **Cost-Complexity Pruning** and **Reduced Error Pruning** with **Cross-Validation**? *1 point*

- [x] In Reduced Error Pruning, the pruned tree error can never be less than the original tree on the training dataset.
- [x] In Cost Complexity Pruning, the pruned tree error can never be less than the original tree on the training dataset.
- [ ] In Reduced Error Pruning, the pruned tree error can never be less than the original tree on the validation dataset.
- [ ] In Cost Complexity Pruning, the pruned tree error can never be less than the original tree on the validation dataset.

Yes, the answer is correct.
Score: 1

Accepted Answers:
In Cost Complexity Pruning, the pruned tree error can never be less than the original tree on the training dataset.
In Reduced Error Pruning, the pruned tree error can never be less than the original tree on the training dataset.

2) Which of the following classifiers would you use if your boss didn't have faith in Machine Learning algorithms and required you to explain how the classifier makes it's predictions? *2 points*

- ○ Regression Trees (with Linear hyperplane in each region)
- ○ Linear Regression
- ◉ Regression Trees (with Constant in each region)
- ○ Neural Networks

Yes, the answer is correct.
Score: 2

Accepted Answers:
Regression Trees (with Constant in each region)

3) *1 point*

Price	Maintenance	Capacity	Airbag	Profitable?
low	low	2	no	yes
low	med	4	yes	no
low	low	4	no	yes
low	high	4	no	no
med	med	4	no	no
med	med	4	yes	yes
med	high	2	yes	no
med	high	5	no	yes
high	med	4	yes	yes
high	high	2	yes	no
high	high	5	yes	yes

In the table given above, what would be the best parameter to split on (at the root), given that we use the Gini-Index measure?

- ○ Airbag
- ○ Maintenance
- ○ Price
- ○ Capacity

No, the answer is incorrect.
Score: 0

Accepted Answers:
Maintenance

4) For the above dataset, which of the following splits is the best when the Gini-Index is used? Assume that the split is *binary.* ***1 point***

- price - {low, med}|{high}
- capacity - {2}|{4, 5}
- maintenance - {high}|{med, low}
- maintenance - {high, med}|{low}

No, the answer is incorrect.
Score: 0

Accepted Answers:
maintenance - {high, med}|{low}

5) In the dataset given above, which of the following would be the best choice given that we use the ***misclassification error*** as the metric to split on? Also assume that we use *binary* splits. ***0 points***

- price - {low, med}|{high}
- maintenance - {high}|{med, low}
- maintenance - {high, med}|{low}
- capacity - {2}|{4, 5}

No, the answer is incorrect.
Score: 0

Accepted Answers:
maintenance - {high}|{med, low}
capacity - {2}|{4, 5}

6) Assume that we constructed the following decision tree for the above data: ***0 points***

Now given that we get an input record that has missing values for *maintenance* and *price.*
We decide to use the ***Fragment*** method of handling missing values.

What are the percentages of *yes* and *no* that we obtain?

- 75% Yes, 25% No
- 67.5% Yes, 32.5% No
- 50% Yes, 50% No
- 32.5% Yes, 67.5% No

No, the answer is incorrect.
Score: 0

Accepted Answers:
67.5% Yes, 32.5% No
32.5% Yes, 67.5% No
50% Yes, 50% No
75% Yes, 25% No

7) In which of the following situations is it appropriate to introduce a new category 'Missing' for missing values? ***2 points***

- [x] When values are missing because the 108 emergency operator is sometimes attending a very urgent distress call.
- [ ] When values are missing because the attendant spilled coffee on the papers from which the data was extracted.
- [ ] When values are missing because the warehouse storing the paper records went up in flames and burnt parts of it.
- [x] When values are missing because the nurse/doctor finds the patient's situation too urgent.

Yes, the answer is correct.
Score: 2

Accepted Answers:
When values are missing because the 108 emergency operator is sometimes attending a very urgent distress call.
When values are missing because the nurse/doctor finds the patient's situation too urgent.

8) An important factor that influences the variance of decision trees is the average height of the tree. For the same dataset, if we limited the height of the trees to some *H,* how would the variance of the decision tree algorithm be affected? ***2 points***

- (•) Variance increases with tree length *H.*
- Variance decreases with tree length *H.*
- Variance is unaffected by tree length *H.*

Yes, the answer is correct.
Score: 2

Accepted Answers:
Variance increases with tree length H.

9) For the ROC curve of True positive rate vs False positive rate, which of the following are true? ***2 points***

- ○ The curve is always concave (negative convex).
- ○ The curve is never concave.
- ◉ The curve may or may not be concave

Yes, the answer is correct.
Score: 2

Accepted Answers:
The curve may or may not be concave

10) In a decision tree, if we decide to swap out the usual splits (of the form $x_i < k$ or $x_i > k$) and instead used a linear combination of features instead, (like $\beta^T x + \beta_0$), where the parameters of the hyperplane β, β_0 are also simultaneously learnt, which of the following statements would be **true**? ***3 points***

- ☑ If we trained only a single step of the decision tree (only the root), the system is equivalent to a perceptron.
- ☐ If we trained only a single step of the decision tree (only the root), the system is equivalent to an SVM.
- ☑ The resulting system cannot solve the XOR problem (refer to the 'Perceptron' lectures)
- ☐ The resulting system can theoretically reach 100% accuracy on the training data set.

No, the answer is incorrect.
Score: 0

Accepted Answers:
If we trained only a single step of the decision tree (only the root), the system is equivalent to a perceptron.
The resulting system can theoretically reach 100% accuracy on the training data set.

Previous Page End

Unit 15 - Hypothesis Testing

Quiz #7

The due date for submitting this assignment has passed.
As per our records you have not submitted this assignment.

Due on 2017-03-16, 23:59 IST.

1) You have an oracle that provides you with N random p-dimensional data points everytime you query it. ***2 points***
You decide to write a program to compute the mean(the average, actually) of the p-dimensional distribution that the oracle is drawing the points from. To test if your program is working, you make a few queries and notice that you're getting N unique points as input, each time, from the oracle. Based on your success, you decide to run your program. At this point, the oracle decides to give you a fixed set of N points for all your successive queries.

Let $\bar{X}_{old}$ be the average reported by your program during testing and $\bar{X}_{new}$ be the average reported by your program after deployment. Which of the following are true about the sampling distribution for $\bar{X}_{old}$?

- [ ] It converges to a Gaussian distribution as $N \to \infty$.
- [ ] It is an uniform distribution.
- [ ] The distribution consists of only a single point. All other points have zero probability of occurrence.
- [ ] It has (N-1) degrees of freedom.

No, the answer is incorrect.
Score: 0

Accepted Answers:
It converges to a Gaussian distribution as $N \to \infty$.
It has (N-1) degrees of freedom.

2) Now, what can you say about the sampling distribution for $\bar{X}_{new}$? ***2 points***

- [ ] It converges to a Gaussian distribution as $N \to \infty$.
- [ ] It is an uniform distribution.
- [ ] The distribution consists of only a single point. All other points have zero probability of occurrence.
- [ ] It has (N-1) degrees of freedom.

No, the answer is incorrect.
Score: 0

Accepted Answers:
The distribution consists of only a single point. All other points have zero probability of occurrence.
It has (N-1) degrees of freedom.

3) You are an administrator for transportation controllers. Part of your job description involves supervising ***3 points***
algorithms that set traffic lights based on vehicle density, movement, etc. Currently, you are employing Algorithm A and have been using it for 5 years now. You've recently come across Algorithm B and decide to test if B is better than A. Performance is measured in the average fuel economy of cars on the road. A higher value indicates lesser time spent waiting in traffic and vice-versa. From 5 years of data(collected on a daily basis), you find out that $\mu_A = 12$ kmpl and $\sigma_A = 2.5$ kmpl (mean and variance of fuel economy). You do a test run of B for **25 days** and find that $\bar{X}_B = 13.0$ kmpl. Based on these numbers, you formulate two hypotheses.
$H_0 : \mu_A = \mu_B H_1 : \mu_A < \mu_B$
What is the z-value obtained during your Hypothesis Testing?

- ( ) 0.02523
- ( ) 0.0228
- ( ) 2.00
- ( ) 0.50

No, the answer is incorrect.
Score: 0

Accepted Answers:
2.00

4) What is your conclusion if you had decided on using a p-value of 0.01 before you began the hypothesis test? ***1 point***

- Accept the Null hypothesis; Reject the Alternate hypothesis.
- Reject the Null hypothesis; Accept the Alternate hypothesis.
- Accept both hypotheses.
- Reject both hypotheses due to lack of information.

No, the answer is incorrect.
Score: 0

Accepted Answers:
Accept the Null hypothesis; Reject the Alternate hypothesis.

5) Now, say, you modify Algorithm B to use a different sub-procedure that might affect it's performance. ***2 points***
Let's call this Algorithm B-1. Running Algorithm B-1 under same, but simulated, conditions as Algorithm B produces $\bar{X}_{B-1} = 13.5$ kmpl.
What is the p-value returned by your z-test upon using Algorithm B-1 in place of B, keeping the hypotheses same?

- 0.013
- 0.0056
- 0.056
- 0.0013

No, the answer is incorrect.
Score: 0

Accepted Answers:
0.0013

6) You are the owner of a casino in Las Vegas. For years, you've been secretly cheating your customers ***3 points***
by using a biased coin that would return heads with a probability of 0.65 (based on your records for the last 12 years), instead of the usual 0.50. One of your rookie employees has come up with a new coin design that is very visually appealing. You are, however, skeptical that this coin might become more fair than the original one and cost you some of your profits. Being a Statistics Major, you decide to use simple Hypothesis Testing to conclude if your new coin will cost you money or not. Your rookie employee tosses the coin 10 times and the following are the results.
T,H,H,T,T,T,H,T,H,H
You formulate the following two hypotheses.
Null : New Coin doesn't cost you any profits
Alternate : New Coin will cost you profits
You are prepared to reject the Null hypothesis in favour of the Alternate hypothesis if your p-value is less than 0.01.
What p-value do you obtain from your statistical test?

- 0.0075
- 0.0025
- 0.0050
- 0.0057

No, the answer is incorrect.
Score: 0

Accepted Answers:
0.0075

7) Now, you remember than the most commonly used Maximum Likelihood estimator for **sample** variance ***2 points***
estimation is actually **biased** and that an **unbiased** sample variance estimation can be done using the following formula.

$$s^2_{unbiased} = \sqrt{\frac{\sum_{i=0}^{i=N-1}(X_i - \bar{X})^2}{N-1}}$$

What is the p-value you obtain using the unbiased variance estimator?

- 0.0075
- 0.0096
- 0.0104
- 0.0125

No, the answer is incorrect.
Score: 0

Accepted Answers:
0.0096

8) You are an algorithms engineer at Okkulus Inc, a Virtual Reality company. You are responsible for ***0 points***
writing algorithms that allow seamless, low-latency communication between the controller and the processor.
You have been keeping up with the research community and you have narrowed down on two algorithms A and B.
You implement both and measure their latencies for the same benchmark movements (in order), in a simulated environment and under the same conditions.
The following are the latencies, in milliseconds(ms).
A - 65, 40, 71, 77, 49, 52, 72
B - 62, 49, 62, 81, 49, 71, 56

Based on exploratory analysis, you hypothesize that algorithm A has lower latencies than algorithm B. You formulate the following hypotheses.
Null : Both algorithms have the same mean latencies
Alternate : Mean latency of algorithm A is lower than that of algorithm B

You decide to ditch your Null hypothesis in favour of your Alternate hypothesis if the resulting p-value is less than 0.05.
What is the p-value obtained during your statistical test?

- 0.04737
- 0.004737
- 0.3727
- 0.03727

No, the answer is incorrect.
Score: 0

Accepted Answers:
0.3727

9) Which of the following definitions of Confidence Intervals are correct for statistics calculated on samples of size **N**? ***1 point***
Note : ϵ is the total width of the confidence interval; C is the confidence in percentage.

- $\bar{X} \in [\mu - \frac{\epsilon}{2}, \mu + \frac{\epsilon}{2}]$ for at least C % of samples of size N drawn from the underlying distribution.
- $\mu \in [\bar{X} - \frac{\epsilon}{2}, \bar{X} + \frac{\epsilon}{2}]$ for at least C% of samples of size N drawn from the underlying distribution.
- $\bar{X} \in [\mu - \epsilon, \mu + \epsilon]$ for at at most (100-C)% of samples of size N drawn from the underlying distribution.
- $\mu \in [\bar{X} - \epsilon, \bar{X} + \epsilon]$ for at most (100-C)% of samples of size N drawn from the underlying distribution.

No, the answer is incorrect.
Score: 0

Accepted Answers:
$\mu \in [\bar{X} - \frac{\epsilon}{2}, \bar{X} + \frac{\epsilon}{2}]$ for at least C% of samples of size N drawn from the underlying distribution.

10)Which of the following can reduce the width of Confidence Intervals? ***1 point***
Note - C is the confidence in percentage, N is the sample size

- [ ] Increase C
- [ ] Reduce C
- [ ] Increase N
- [ ] Reduce N

No, the answer is incorrect.
Score: 0

Accepted Answers:
Reduce C
Increase N

Previous Page

End

Unit 16 - Ensemble Methods

Quiz #8

The due date for submitting this assignment has passed.
As per our records you have not submitted this assignment.

Due on 2017-03-22, 23:59 IST.

1) How does bagging help in designing better classifiers? *1 point*

- [ ] It helps reduce variance
- [ ] It helps reduce bias
- [ ] If the parameters of the resultant classifiers are fully uncorrelated (independent), then bagging is inefficient.
- [ ] If the parameters of the resultant classifiers are fully correlated, then bagging is inefficient.

No, the answer is incorrect.
Score: 0

Accepted Answers:
It helps reduce variance
If the parameters of the resultant classifiers are fully correlated, then bagging is inefficient.

2) If you have a *bad* classifier, which of the following ensemble methods will give the worst performance when including the given classifier? *1 point*

- ○ AdaBoost
- ○ Gradient Boosting
- ○ Bagging
- ○ Committee Machine

No, the answer is incorrect.
Score: 0

Accepted Answers:
Bagging

3) Which among the following prevents overfitting when we perform bagging? *1 point*

- ○ The use of sampling with replacement as the sampling technique
- ○ The use of weak classifiers
- ○ The use of classification algorithms which are not prone to overfitting
- ○ The practice of validation performed on every classifier trained

No, the answer is incorrect.
Score: 0

Accepted Answers:
The use of sampling with replacement as the sampling technique

4) Considering the loss functions used by which one among Gradient Boosting and AdaBoost is less susceptible to outliers? *2 points*

- ○ AdaBoost
- ○ Gradient Boost
- ○ On average, both are equally susceptible.

No, the answer is incorrect.
Score: 0

Accepted Answers:
Gradient Boost

5) Which of the following statements are TRUE when comparing **Committee Machines** and **Stacking** *2 points*

- [ ] Committee Machines are, in general, special cases of 2-layer stacking where the second-layer classifier provides uniform weightage.
- [ ] Both Committee Machines and Stacking have similar mechanisms, but Stacking uses different classifiers while Committee Machines use similar classifiers.
- [ ] Committee Machines are more powerful than Stacking
- [ ] Committee Machines are less powerful than Stacking

No, the answer is incorrect.
Score: 0

Accepted Answers:
Committee Machines are less powerful than Stacking
Committee Machines are, in general, special cases of 2-layer stacking where the second-layer classifier provides uniform weightage.

6) In terms of extent of simultaneous execution (parallel ability), which of the following methods are the least suited for simultaneous execution? ***2 points***

- Bagging
- Random Forests
- Boosting
- Stacking

No, the answer is incorrect.
Score: 0

Accepted Answers:
Boosting

7) Boosting techniques typically give very high accuracy classifiers by sequentially training a collection of similar low-accuracy classifiers. ***2 points***
Which of the following statements are true with respect to Boosting?

- [ ] LogitBoost (like AdaBoost, but with Logistic Loss instead of Exponential Loss) is less susceptible to overfitting than AdaBoost.
- [ ] Boosting techniques tend to have low bias and high variance
- [ ] Boosting techniques tend to have low variance and high bias
- [ ] For basic linear regression classifiers, there is no effect of using Gradient Boosting.

No, the answer is incorrect.
Score: 0

Accepted Answers:
LogitBoost (like AdaBoost, but with Logistic Loss instead of Exponential Loss) is less susceptible to overfitting than AdaBoost.
Boosting techniques tend to have low bias and high variance
For basic linear regression classifiers, there is no effect of using Gradient Boosting.

8) While using Random Forests, if the input data is such that it contains a large number (>80%) of irrelevant features (the target variable is independent of the these features), which of the following statements are TRUE? ***3 points***

- [ ] Random forests have *increased* performance as the *number* of irrelevant features increases.
- [ ] Random Forests have reduced performance as the *number* of irrelevant features increases
- [ ] Random Forests have reduced performance as the *fraction* of irrelevant features increases.

No, the answer is incorrect.
Score: 0

Accepted Answers:
Random Forests have reduced performance as the fraction of irrelevant features increases.

Previous Page
End

Unit 17 - Graphical Models

Quiz #9

The due date for submitting this assignment has passed.
As per our records you have not submitted this assignment.

Due on 2017-03-29, 23:59 IST.

This Quiz consists of 10 questions, all based on Week #9 course content - Graphical Models.

You may require the use of a calculator while solving this quiz.

1) Which of the following issues with the Gaussian Distribution are a concern for it being used as Conditional Probability Distributions (CPDs) in Bayesian Networks? ***1 point***

- ☐ A probability distribution cannot be explained just by using it's mean and variance.
- ☐ The Gaussian distribution is a unimodal distribution. It cannot capture multi-modal distributions.
- ☐ It's cumulative distribution function cannot be computed analytically.
- ☐ There is no multi-variate Gaussian form to model distributions of the form $Pr(A \mid B_1, B_2 \ldots B_N)$.

No, the answer is incorrect.
Score: 0

Accepted Answers:
A probability distribution cannot be explained just by using it's mean and variance.
The Gaussian distribution is a unimodal distribution. It cannot capture multi-modal distributions.

2) Select the correct pair of graphs and their implied independence results. ***2 points***
Note : Text in bold is to be considered as a single node; Arrows represent dependence as in normal Bayesian Networks.

- ☐ $\mathbf{A} \rightarrow \mathbf{C} \leftarrow \mathbf{B}$ -> **A** is independent of **B** given **C**
- ☐ $\mathbf{A} \rightarrow \mathbf{C} \leftarrow \mathbf{B}$ -> **A** depends on **B** if **C** is known.
- ☐ $\mathbf{A} \rightarrow \mathbf{C} \rightarrow \mathbf{B}$ -> **B** is independent of **A** if **C** is known.
- ☐ $\mathbf{A} \leftarrow \mathbf{C} \rightarrow \mathbf{B}$ -> **A** is independent of **B** given **C**.

No, the answer is incorrect.
Score: 0

Accepted Answers:
*$A \rightarrow C \leftarrow B$ -> **A** depends on **B** if **C** is known.*
*$A \rightarrow C \rightarrow B$ -> **B** is independent of **A** if **C** is known.*
*$A \leftarrow C \rightarrow B$ -> **A** is independent of **B** given **C**.*

3) Which of the following statements are true regarding Bayesian Networks and Markov Random Fields? ***2 points***

- ☐ Bayesian Networks has one Conditional Probability Distribution for each directed edge.
- ☐ For each edge in a Markov Random Field, there exists a factor/potential function.
- ☐ A Bayesian Network can have a **node prior** to add some probability mass to the node regardless of the CPDs.
- ☐ Inference in Bayesian Networks is harder than Markov Random Fields.
- ☐ Inference in Markov Random Fields is harder than Bayesian Networks.
- ☐ A Bayesian Network permits loops in its graph description.
- ☐ A Markov Random Field permits loops in its graph description.

No, the answer is incorrect.
Score: 0

Accepted Answers:
Bayesian Networks has one Conditional Probability Distribution for each directed edge.
Inference in Markov Random Fields is harder than Bayesian Networks.
A Markov Random Field permits loops in its graph description.

4) Given the following Bayesian Network, select the right set of independence rules implied by the graph. ***1 point***

- [ ] A is independent of B given C
- [ ] B is independent of C given D
- [ ] C is independent of D if E is not known
- [ ] A is independent of D if C is known

No, the answer is incorrect.
Score: 0

Accepted Answers:
C is independent of D if E is not known

5) Using the following CPDs and probabilities, compute the probability of the assignment (0,1,1,0,1) to the variables (A,B,C,D,E). ***2 points***

Pr(C=0|A,B) for (A,B) =

- (0,0) = 0.2
- (0,1) = 0.4
- (1,0) = 0.6
- (1,1) = 0.8

Pr(E=0|C,D) for (C,D) =

- (0,0) = 0.1
- (0,1) = 0.3
- (1,0) = 0.5
- (1,1) = 0.7

Pr(A=0) = 0.4

Pr(B=0) = 0.6

Pr(D=0) = 0.9

- ○ 0.4328
- ○ 0.0432
- ○ 0.1008
- ○ 0.00966

No, the answer is incorrect.
Score: 0

Accepted Answers:
0.0432

6) Using the above probabilities, what is the probability of the assignment (1,0,1,0) to the variables (A,B,C,E) ? ***2 points***

- ○ 0.07488
- ○ 0.01668
- ○ 0.10648
- ○ 0.20116

No, the answer is incorrect.
Score: 0

Accepted Answers:
0.07488

7) Given the assignment (1,0,1,1) to the variables (B,C,D,E), what is the probability that A is 1? ***2 points***

- ○ 0.0108
- ○ 0.25
- ○ 0.75
- ○ 0.9892

No, the answer is incorrect.
Score: 0

Accepted Answers:
0.75

8) Consider the following Markov Random Field. ***2 points***

Which of the following nodes will have no effect on D given the Markov Blanket of D?

- [ ] A
- [ ] B
- [ ] C
- [ ] E
- [ ] F

- [ ] G
- [ ] H
- [ ] I
- [ ] J

No, the answer is incorrect.
Score: 0

Accepted Answers:
A
I
J

9) Which of the following statement are true w.r.t Hidden Markov Models (HMMs) ? ***2 points***

- [ ] One can obtain probability of observing a sequence using HMMs.
- [ ] HMMs assume Markov property over y_i: $Pr(y_i|y_{i-1},y_{i-2}\ldots y_0) = Pr(y_i|y_{i-1})$
- [ ] HMMs assume Markov Property over x_i: $Pr(x_i|x_{i-1},x_{i-2}\ldots x_0) = Pr(x_i|x_{i-1})$.
- [ ] One can use the Viterbi Algorithm to obtain the MAP sequence of x_i's given y_i's.
- [ ] One can use the Viterbi Algorithm to obtain the MAP sequence of y_i's given x_i's.

No, the answer is incorrect.
Score: 0

Accepted Answers:
One can obtain probability of observing a sequence using HMMs.
HMMs assume Markov Property over x_i: $Pr(x_i|x_{i-1},x_{i-2}\ldots x_0) = Pr(x_i|x_{i-1})$.
One can use the Viterbi Algorithm to obtain the MAP sequence of x_i*'s given* y_i*'s.*

10) Select the correct pairs of (Graphical Model, Inference Algorithm) ***2 points***

- [ ] (Bayesian Networks, Variable Elimination)
- [ ] (Viterbi Algorithm, Markov Random Fields)
- [ ] (Viterbi Algorithm, Hidden Markov Models)
- [ ] (Belief Propagation, Markov Random Fields)
- [ ] (Variable Elimination, Markov Random Fields)

No, the answer is incorrect.
Score: 0

Accepted Answers:
(Bayesian Networks, Variable Elimination)
(Viterbi Algorithm, Hidden Markov Models)
(Belief Propagation, Markov Random Fields)
(Variable Elimination, Markov Random Fields)

Previous Page

End

Unit 18 - Clustering

Quiz #10

The due date for submitting this assignment has passed.
As per our records you have not submitted this assignment.

Due on 2017-04-06, 23:59 IST.

1) In the given distribution of data points (3 classes denoted by their respective colors), which of the following algorithms performs the best? ***1 point***

- K-means
- DBSCAN
- Single-link Hierarchical Clustering
- K-mediods

No, the answer is incorrect.
Score: 0

Accepted Answers:
DBSCAN

2) In the given distribution of data points (3 classes denoted by their respective colours), if we used DBSCAN, which of the following epsilon values works best? ***1 point***

- 150.0
- 2.0
- 20.5
- 0.25

No, the answer is incorrect.
Score: 0

Accepted Answers:
2.0

3) Which of the following statements are true about K-means? ***1 point***

- [ ] The number of clusters is given.
- [ ] All clusters are of the same shape
- [ ] All the clusters have the same number of points.

No, the answer is incorrect.
Score: 0

Accepted Answers:
The number of clusters is given.
All clusters are of the same shape

4) Considering single-link and complete-link hierarchical clustering, is it possible for a point to be closer to points in other clusters than to points in its own cluster? If so, in which approach will this tend to be observed? ***2 points***

- [ ] Single-link
- [ ] Complete-link
- [ ] Centroid-based
- [ ] None of the above

No, the answer is incorrect.
Score: 0

Accepted Answers:
Complete-link
Centroid-based

5) In the lecture on the BIRCH algorithm, it is stated that using the number of points ***N***, sum of points ***SUM*** and sum of squared points ***SS,*** we can determine the centroid and radius of the combination of any two clusters **A** and **B**. ***2 points***
How do you determine the centroid of the combined cluster? (In terms of **N,SUM & SS** of both the clusters)

- $SUM_a + SUM_b$

$\frac{SUM_a}{N_a} + \frac{SUM_b}{N_b}$

$\frac{SUM_a + SUM_b}{N_a + N_b}$

$\frac{SS_a + SS_b}{N_a + N_b}$

No, the answer is incorrect.
Score: 0

Accepted Answers:
$\frac{SUM_a + SUM_b}{N_a + N_b}$

6) If your supervisor gives you a machine with 2TB of RAM and asks you to find clusters in a dataset of roughly 5 billion data points (each data point taking roughly 128 bytes of data). Considering the large data size, which of the following algorithms would you use to efficiently find patterns in the data? ***0 points***

- BIRCH
- CURE
- K-means
- Logistic Regression

No, the answer is incorrect.
Score: 0

Accepted Answers:
BIRCH
CURE

7) Consider the following criteria for clustering with threshold graphs (hierarchical clustering): ***3 points***

A ***p***-complete cluster is one where the ratio $\frac{E}{\frac{V(V-1)}{2}}$ is greater than ***p***. This quantity is the fraction of edges present in the cluster out of all possible edges within the cluster.

If we define ***p***-complete clusters of the thresholded graphs as the clustering algorithm, is it well-defined?

- No, it is not well-defined.
- Yes, it is well-defined.

No, the answer is incorrect.
Score: 0

Accepted Answers:
Yes, it is well-defined.

Previous Page
Next Page

Unit 18 - Clustering

PA#2

The due date for submitting this assignment has passed.
As per our records you have not submitted this assignment.

Due on 2017-04-10, 23:59 IST.

The following questions need to be answered using Weka's Clustering Algorithms.

If you haven't already downloaded & installed Weka, please find the instructions here for all 3 platforms (OS X, Windows & Linux):
http://web.engr.oregonstate.edu/~tgd/classes/geo599/diabetes/java-and-weka.html

Here is a document that outlines the procedure to run a K-means algorithm on the **Iris** dataset:
http://modelai.gettysburg.edu/2016/kmeans/assets/iris/Clustering_Iris_Data_with_Weka.pdf

The dataset ARRF files for the 3 datasets can be found here:

Iris: https://onlinecourses.nptel.ac.in/noc17_cs17/assets/img/iris.arff
Path-based:https://onlinecourses.nptel.ac.in/noc17_cs17/assets/img/pathbased.arff
Spiral: https://onlinecourses.nptel.ac.in/noc17_cs17/assets/img/spiral.arff

Special Notes: If the DBSCAN Algorithm is not available in the 'Choose' menu of the 'Cluster' tab, you need to quit the Explorer, go to 'Tools->Package Manager' and install the 'optics-dbscan' package.

1) With the IRIS dataset, run K-means ('SimpleKMeans') with K=2. What is the percentage of points incorrectly clustered? ***1 point***

- 33%
- 50%
- 39%
- 11%

No, the answer is incorrect.
Score: 0

Accepted Answers:
33%

2) On the same IRIS dataset, run Hierarchical Clustering in both Complete-link and Single-link modes. Report the percentage of wrongly clustered data. (Stop at exactly 3 clusters in both cases) ***1 point***

- 34%, 15%
- 12%, 34%
- 34%, 12%
- 15%, 34%

No, the answer is incorrect.
Score: 0

Accepted Answers:
12%, 34%

3) On which of the following clusters does Single link Hierarchical Clustering perform well? ***1 point***

- Path-based
- Spiral
- Iris

No, the answer is incorrect.
Score: 0

Accepted Answers:
Spiral

4) In the PathBased dataset, use Hierarchical Clustering and try all the modes available. ***2 points***

Which one gives the best clustering performance?
Visualize each case and try to intuitively understand why this happens.

- Single-link
- Complete-link
- Average-link
- Mean
- Centroid

No, the answer is incorrect.
Score: 0

Accepted Answers:
Centroid

5) DBSCAN Calibration: ***2 points***

DBSCAN is an algorithm that generally needs a significant amount of parameter tuning to get a non-trivial result. An epsilon value too high will cause the entire dataset to be placed into a single cluster and an epsilon value too small will cause a very large number of small clusters.
Which of the following options best describes the epsilon values that give a reasonable clustering?

Note: You will have to visualize the results for various epsilon values to understand this.

- 0.04 - 0.09
- 0.4 - 0.9
- 0.01 - 0.04
- 1 - 4

No, the answer is incorrect.
Score: 0

Accepted Answers:
0.04 - 0.09
0.4 - 0.9
0.01 - 0.04
1 - 4

Unit 21 - Learning Theory

Quiz

The due date for submitting this assignment has passed.
As per our records you have not submitted this assignment.

Due on 2017-04-12, 23:59 IST.

This quiz consists of 9 questions on the following topics.

- Gaussian Mixture Models & Expectation Maximization
- Spectral Clustering
- Learning Theory

1) In Gaussian Mixture Models, π_i are the Mixing coefficients. Select the **correct** conditions that the mixing coefficients need to satisfy for a valid GMM model. ***2 points***

- ☐ $0 \leq \pi_i \leq 1 \quad \forall i$
- ☐ $-1 \leq \pi_i \leq 1 \quad \forall i$
- ☐ $\sum_i \pi_i = 1$
- ☐ $\sum_i \pi_i$ need not be bounded.

No, the answer is incorrect.
Score: 0

Accepted Answers:
$0 \leq \pi_i \leq 1 \quad \forall i$
$\sum_i \pi_i = 1$

2) During parameter estimation for a GMM model using data **X**, which of the following quantities are you minimizing (directly or indirectly)? ***1 point***

- ○ Log-likelihood
- ○ Negative Log-likelihood (NLL)
- ○ Cross-entropy
- ○ Residual Sum of Squares (RSS)

No, the answer is incorrect.
Score: 0

Accepted Answers:
Negative Log-likelihood (NLL)

3) Expectation-Maximization, or the EM algorithm, consists of two steps - E step and the M-step. Using the following notation, select the correct set of equations used at each step of the algorithm. ***2 points***

Notation.
$\mathbf{X}$ – Known/Given variables/data
$\mathbf{Z}$ – Hidden/Unknown variables
θ – Total set of parameters to be learned
θ_k – Values of all the parameters after stage **k**
$\mathbf{Q}(\cdot, \cdot)$ – The Q-function as described in the lectures

- ☐ $\mathbf{E} - \mathrm{E}_{\mathbf{Z}|\mathbf{X},\theta_{m-1}}[log(Pr(\mathbf{X},\mathbf{Z}|\theta))]$
- ☐ $\mathbf{E} - \mathrm{E}_{\mathbf{Z}|\mathbf{X},\theta}[log(Pr(\mathbf{X},\mathbf{Z}|\theta_m))]$
- ☐ $\mathbf{M} - argmax_\theta \sum_{\mathbf{Z}} Pr(\mathbf{Z}|\mathbf{X},\theta_{m-1}) \cdot log(Pr(\mathbf{X},\mathbf{Z}|\theta))$
- ☐ $\mathbf{M} - argmax_\theta \mathbf{Q}(\theta, \theta_{m-1})$
- ☐ $\mathbf{M} - argmax_\theta \mathbf{Q}(\theta, \theta_{m-2})$

No, the answer is incorrect.

Miscellaneous

Score: 0

Accepted Answers:

$E - E_{Z|X,\theta_{m-1}}[log(Pr(X,Z|\theta))]$

$M - argmax_\theta \sum_Z Pr(Z|X,\theta_{m-1}) \cdot log(Pr(X,Z|\theta))$

$M - argmax_\theta Q(\theta,\theta_{m-1})$

4) Using notation as in the Spectral Clustering lectures, select the correct equations involving $\vec{s}$ (using vector notation for column vectors) ***2 points***

- ☐ $||\vec{s}||_2 = 1$
- ☐ $||\vec{s}||_2 = \sqrt{n}$
- ☐ $\vec{s} \cdot \vec{s} = n$
- ☐ $s_i \in \{-1, +1\} \quad \forall i$
- ☐ $s_i \in [0,1] \quad \forall i$

No, the answer is incorrect.
Score: 0

Accepted Answers:

$||\vec{s}||_2 = \sqrt{n}$

$\vec{s} \cdot \vec{s} = n$

$s_i \in \{-1, +1\} \quad \forall i$

5) Select all the correct statements regarding the Unnormalized Laplacian, as discussed in the lectures. ***3 points***

- ☐ 1 is an eigen-vector of **L**
- ☐ 0 is an eigen-value corresponding to the first eigen-vector of **L**.
- ☐ When ordered in increasing value of eigen-values, $\vec{v}_1$, the eigen-vector corresponding to the smallest eigen-value leads to a degenerate solution for the Spectral Clustering problem.
- ☐ $\forall i, \quad \sum_j \mathbf{L}_{ij} = 0$
- ☐ $\forall j, \quad \sum_i \mathbf{L}_{ij} = 0$

No, the answer is incorrect.
Score: 0

Accepted Answers:

1 is an eigen-vector of ***L***

0 is an eigen-value corresponding to the first eigen-vector of ***L****.*

When ordered in increasing value of eigen-values, $\vec{v}_1$*, the eigen-vector corresponding to the smallest eigen-value leads to a degenerate solution for the Spectral Clustering problem.*

$\forall i, \quad \sum_j \boldsymbol{L}_{ij} = 0$

$\forall j, \quad \sum_i \boldsymbol{L}_{ij} = 0$

6) Let's order the eigen-values of **L** in increasing order of their value. This also induces an implicit ordering on the eigen-vectors. Consider, for the purpose of this question, that if **L** has multiple eigen-vectors for the same eigen-value, we only pick a *representative* eigen-vector. Let the number of eigen-vectors now be **N (**and **N** > 3**)**. Which of the following eigen-vectors is the Fiedler-vector? ***2 points***

- ☐ $\vec{v}_1$
- ☐ $\vec{v}_n$
- ☐ $\vec{v}_2$
- ☐ $\vec{v}_{n-1}$
- ☐ $\vec{v}_k$ such that the eigen-value λ_k is the first non-zero eigen-value in the ordering.

$\vec{v}_k$ such that the eigen-value λ_k is the last non-zero eigen-value in the ordering.

No, the answer is incorrect.
Score: 0

Accepted Answers:

$\vec{v}_2$

$\vec{v}_k$ such that the eigen-value λ_k is the first non-zero eigen-value in the ordering.

7) You want to toss a **fair** coin a number of times and obtain the probability of it falling on it's heads by taking a simple average. What is the estimated number of times you'll have to toss the coin to make sure that your estimated probability is **within 10%** of the actual probability, at least **90% of the time**? ***2 points***

- 400*ln(20)
- 800*ln(20)
- 200*ln(20)
- 100*ln(40)

No, the answer is incorrect.
Score: 0

Accepted Answers:
*200*ln(20)*

8) You're considering a new set of regressors where there are only **100** different parameters settings possible with the architecture you have set up. What is the minimum number of data points you need to use to have an error not more than **0.5** away from the optimal parameter setting with **at least 95%** confidence? ***3 points***

- 2*ln(400)
- 4*ln(4000)
- 2*ln(4000)
- 8*ln(2000)

No, the answer is incorrect.
Score: 0

Accepted Answers:
*2*ln(4000)*

9) You're using a Neural Network model that can be specified completely by **K** 32-bit floating point numbers. What is the sample complexity of your Neural Network model for some assignment to other parameters like confidence and absolute error? ***2 points***

- $\mathcal{O}(K)$
- $\mathcal{O}(2^K)$
- $\mathcal{O}(K^2)$
- $\mathcal{O}(\log_2 K)$

No, the answer is incorrect.
Score: 0

Accepted Answers:
$\mathcal{O}(K)$

Previous Page

End

Unit 23 - Reinforcement Learning

Quiz#12

The due date for submitting this assignment has passed.
As per our records you have not submitted this assignment.

Due on 2017-04-21, 23:59 IST.

1) From the point of view of computational performance, in which of the following situations are 1 vs 1 classifiers feasible? ***1 point***

- [ ] Determining the city from which the given piece of text is from. (out of every city in the world)
- [ ] Determining the approximate time of day in a photograph (midday / night / morning / dusk)
- [ ] Determining the species of insect from it's photograph. (From among every possible species known to mankind)
- [ ] Determining whether an animal is an insect, land-mammal or fish from it's picture.

No, the answer is incorrect.
Score: 0

Accepted Answers:
Determining the approximate time of day in a photograph (midday / night / morning / dusk)
Determining whether an animal is an insect, land-mammal or fish from it's picture.

2) State whether the following statement is True or False. ***1 point***
The association rule $A \implies B$ has the same support as the association rule $B \implies A$.

- ○ True
- ○ False

No, the answer is incorrect.
Score: 0

Accepted Answers:
True

3) Use the apriori property to check if the following statement is true: ***1 point***

If an item **I1** occurs only once in an FP-tree, then the sub-tree rooted at **I1** is the prefix tree of all itemsets containing **I1.**

- ○ False
- ○ True

No, the answer is incorrect.
Score: 0

Accepted Answers:
False

4) For a transactional database with 6 elements, **{A,B,C,D,E,F}**, if we were to use the Apriori Algorithm to find the frequent itemsets, then in the *worst case,* how many itemsets would we enumerate? ***0 points***

- ○ 256
- ○ 127
- ○ 255
- ○ 128

No, the answer is incorrect.
Score: 0

Accepted Answers:
255
256
127
128

5) In a Reinforcement Learning problem, if the rewards are stochastic (drawn from a probability distribution) and stationary (the distribution is fixed), which of the following explore-exploit tradeoff strategies work best (qualitatively)? ***1 point***

- ○ Always, 100% Explore

Miscellaneous

- Initially, 50% Explore, 50% Exploit
  After a long time, 99% Explore, 1% Exploit
- Initially, 50% Explore, 50% Exploit
  After a long time, 1% Explore, 99% Exploit
- Always, 100% Exploit

No, the answer is incorrect.
Score: 0

Accepted Answers:
Initially, 50% Explore, 50% Exploit
After a long time, 99% Explore, 1% Exploit

6) In a tournament classifier with N classes. What is the complexity of the number of classifiers we require? ***2 points***

- $\mathcal{O}(N^2)$
- $\mathcal{O}(N.log(N))$
- $\mathcal{O}(N)$
- $\mathcal{O}(log(N))$

No, the answer is incorrect.
Score: 0

Accepted Answers:
$\mathcal{O}(N)$

7) In Reinforcement Learning, one method is to maintain a **value** for every state and use various techniques to learn this **value function.** ***2 points***
Which of the following states will likely have the highest **value**, assuming that they all converge to the correct values? (Assume you are playing as **X**)

		X
	O	O
X	O	X
	X	X
		O
O	O	X
		X
		X
X		

No, the answer is incorrect.
Score: 0

Accepted Answers:

	X	*X*
		O
O	*O*	*X*

8) In the context of Reinforcement Learning algorithms, which of the following definitions constitutes a valid Markov State? ***2 points***

- [ ] For **Chess**: Positions of yours and the opponent's remaining pieces
- [ ] For **Tic-Tac-Toe:** A snapshot of the game board (all Xs, Os and empty spaces)
- [ ] For **Poker:** All the cards in the player's hand and the cards
- [ ] For **Chess**: Positions of your pieces and the identities of the opponents defeated pieces.
- [ ] For **Tennis:** Position & Velocity of the ball
- [ ] For **Tennis:** Position of the ball

No, the answer is incorrect.
Score: 0

Accepted Answers:
*For **Chess**: Positions of yours and the opponent's remaining pieces*
*For **Tic-Tac-Toe:** A snapshot of the game board (all Xs, Os and empty spaces)*
*For **Poker:** All the cards in the player's hand and the cards*
*For **Tennis:** Position & Velocity of the ball*

9) Consider two Association Rules **X** and **Y** which represents ***A => B*** and ***C => D*** respectively. If ***lift(X) >*** ***lift(Y)*** but ***conf(X) < conf(Y),*** what can this possibly imply? **2 points**

- ☐ Support of **X** is higher than **Y**
- ☐ The ratio of instances that contain only **B** to the support of **X** is smaller than the ration of instances that contain only **D** to the support of **Y.**
- ☐ **X** is a more useful rule than **Y**
- ☐ **Y** is a more useful rule than **X**

No, the answer is incorrect.
Score: 0

Accepted Answers:
X** is a more useful rule than **Y

10)You are designing a Reinforcement Learning agent for a racing game. Which of the following is theoretically the best reward scheme to ensure the agent learns the true optimal policy? **3 points**

- ○ +5 for reaching the finish line, -1 for going off the road
- ○ +5 for reaching the finish line, -0.1 for every second that passes before the agent reaches the finish line
- ○ +5 for reaching the finish line, -0.1 for every second that passes before the agent reaches the finish line, -1 for the agent going off the road.
- ○ -5 for reaching the finish line, +0.1 for every second that passes before the agent reaches the finish line.

No, the answer is incorrect.
Score: 0

Accepted Answers:
+5 for reaching the finish line, -0.1 for every second that passes before the agent reaches the finish line

Previous Page End