Quiz Submissions - Late Midterm Exam

×

Yuelin Liu (username: yuelin.liu@mail.mcgill.ca)

Attempt 1

Written: Nov 13, 2020 5:53 PM - Nov 13, 2020 10:44 PM

Submission View

Released: Nov 15, 2020 7:00 PM

Question 1 **1 / 1 point**

Please select all option to acknowledge the expectation in regards to academic integrity.

For this exam, I make the following truthful statements:

✓ □
- I will not give any assistance to another student taking this exam or receive assistance from anyone.

✓ □
- I will not copy the content of this exam without instructor's permission.

✓ □
- I understand that acts of academic dishonesty may be penalized to the full extent allowed by the McGill University.

Question 2 **1 / 1 point**

Select the correct ordering of the following ML methods based on labeled data requirements, from most needed labeled data to least

○ dimensionality reduction, classification, imitation learning, reinforcement learning

○ imitation learning, reinforcement learning, classification, dimensionality reduction

○ reinforcement learning, imitation learning, dimensionality reduction, classification

○ classification, dimensionality reduction, reinforcement learning, imitation learning

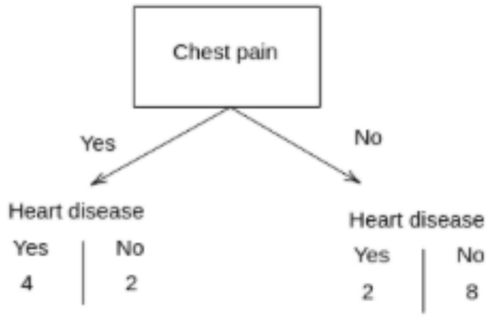
✓ ○ classification, imitation learning, reinforcement learning, dimensionality reduction

▼ Hide Feedback

Reinforcement learning requires a weak type of supervision in the form of reward signal. In comparison dimensionality reduction is completely unsupervised and does not require additional information beyond input features.

Question 3 1 / 1 point

Suppose this one layer decision tree (aka decision stump) classifies the patients into having heart disease or not used on whether they have experienced chest pain. What is the misclassification cost of this tree? (Use 2 decimals format eg, .14)



Answer: 0.25 ✗ **(.25, 1/4, 25%)**

Question 4

1 / 1 point

Tom is running k-NN on 100 data points with 30 features and k=10 with cross validation and getting poor results. Which would you suggest he do to improve his performance?

✓ ○ remove noisy features

○ add more features

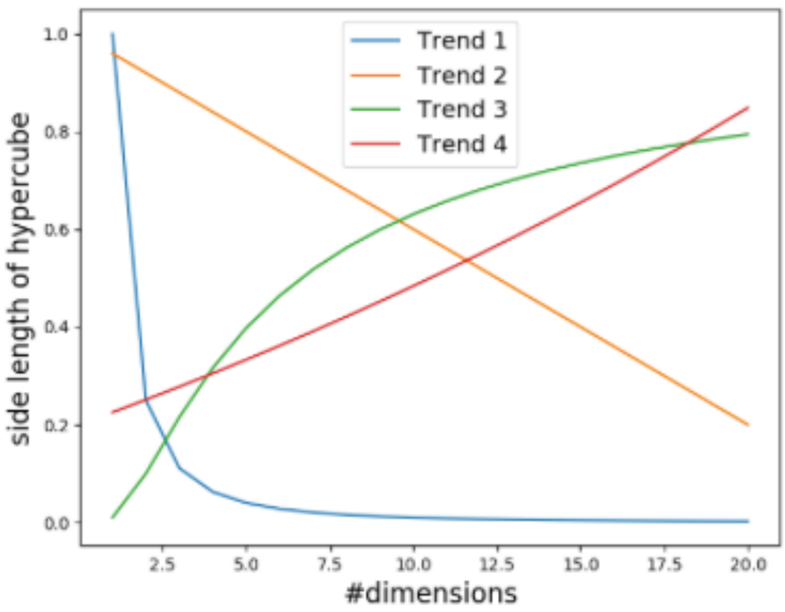
○ remove outliers

○ increase k

Question 5

0 / 1 point

Suppose you have N data points randomly distributed in a D-dimensional hypercube and your model relies on looking at K nearest points of a test point to make its decision. Let us denote the side length of the hypercube around a test point such that it contains K points on average by L. Which of the following plots best represents the relation between L and D?



✗ ○ Trend 1

○ Trend 2

➡ ○ Trend 3

○ Trend 4

▼ Hide Feedback

Volume of hypercube = L^D. This volume should contain k out of points. Hence, L^D ~ (K/N) → L ~ (K/N)^(1/D)

Question 6 **1 / 1 point**

[Numerical] You are performing 5-fold cross-validation to select the best hyperparameter values. You have two hyperparameters and you decide to perform a grid-search over 10 possible values of each hyperparameter (i.e., try all combinations). How many models do you need to train to get a good estimate of the best hyperparameter values?

Answer: 500 ✓

▼ Hide Feedback

100 different combinations of each hyperparameter (10 each). To evaluate each hyperparameter combination, you need to train the model 5 times, once in each fold.

Question 7 **1 / 1 point**

Let θ be the probability of throwing head for a potentially biased coin. Assume a uniform prior over (0,1) for θ. After throwing the coin 5 times we observe 4 heads and 1 tail. What is the probability of the next throw being heads?

○ 4/5

○ 3/4

✓○ 5/7

○ 1/2

▼ Hide Feedback

this is the posterior predictive for Beta-Bernoulli with

$$\beta = \alpha = 1$$

; also called Laplace smoothing.

Question 8 1 / 1 point

Select all the correct statements

✓ □ MAP inference is a compromise between full Bayesian inference and maximum likelihood

✓ □ Maximum likelihood estimation can lead to overfitting

✓ □ Bayesian inference gives a point estimate for model parameters

✓ □ Bayesian inference is often computationally more expensive than the maximum likelihood

Question 9 1 / 1 point

Which of the following statements are true:

✓ □ If p(x1,x2) is Gaussian then p(x1) is Gaussian but p(x1|x2=c) may not be Gaussian in general

✓ □ If Gaussian random variables x1, x2 are not correlated then they are independent

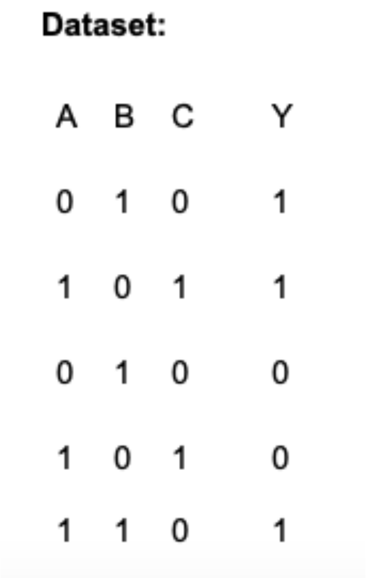
✓ □ If non-Gaussian random variables are correlated then they are not independent

✓ □ The extensive use of Gaussian in statistical models is in part motivated by the central limit theorem

Question 10

1 / 1 point

Consider the following problem where all features (A,B,C) are binary and the label (Y) is binary as well. You are given Dataset 1



Dataset:

A	B	C	Y
0	1	0	1
1	0	1	1
0	1	0	0
1	0	1	0
1	1	0	1

What is P(Y=1|A=0,B=1,C=1) if we use a Bernoulli Naive Bayes. Please give your answer in the form of a fraction b/d where you have **simplified the expression** .

Answer: 8/17 ✓

Question 11

0 / 1 point

Suppose you have a binary classification problem (class y0 vs class y1) with 2 binary features (x0 and x1). You intend to use a Naive Bayes classifier. Your dataset consists of 6 data points (equally distributed between y0 and y1), three of which have the same feature values, i.e. x0 = x1 = 1. However, two of these data points belong to y0 whereas the third data point belongs to y1. After training the Naive Bayes classifier on this dataset, what will its output be for the input: (x0,x1) = (1,1)?

○ Equal probability 0.5 for both classes

○ Data points are not enough to train this Naive Bayes classifier because it has more parameters

➡ ○ Insufficient information. The answer depends on the remaining three data points

✖ ○ Probability ⅔ for class y0 and ⅓ for class y1

▼ Hide Feedback

The answer will depend on the conditional probabilities P(x0=1|y0), P(x1=1|y0) and P(x0=1|y1), P(x1=1|y1). Since Naive Bayes assumes features to be conditionally independent given the class label, the conditional probability estimates depend on the other 3 data points.

Question 12

1 / 1 point

Lets work with EM update with a mixture of two univariate Gaussian distributions.

We will assume that the mean for both the Gaussian components is the same and the standard deviation is different. The latent variable follows the Bernoulli distribution.

$$z \sim \text{Bernoulli}(\theta)$$

$$x|z = \text{k cluster} \sim \mathcal{N}(\mu, \sigma_k)$$

Assume initial values are given as follows.

$$x^{(0)} = 1; x^{(1)} = 3$$

$$\mu = 1; \sigma_0 = 1; \sigma_1 = 2; \theta = 0.3$$

Here θ is the probability of a data point belonging to cluster 1.

Compute the posterior probability for E-step?

$$r^{(0)} = P(z^{(0)} = 1|x^{(0)})$$
$$r^{(1)} = P(z^{(1)} = 1|x^{(1)})$$

For calculations: Use e^(-0.5) = 0.6, e^(-2) = 0.1. Answer by rounding of the final calculation to two decimals. (All the calculations are very simple, they can be done by hand!)

(enter r^(0) in box 1, and r^(1) in box 2)

Answer for blank # 1: 0.18 ✓(50 %)

Answer for blank # 2: 0.56 ✓(50 %)

▼ Hide Feedback

$$
\begin{aligned}
r^{(0)} = P(z^{(0)} = 1|x^{(0)}) &= \frac{\theta\mathcal{N}(x^{(0)}; \mu, \sigma_1)}{\theta\mathcal{N}(x^{(0)}; \mu, \sigma_1) + (1-\theta)\mathcal{N}(x^{(0)}; \mu, \sigma_0)} \\
&= \frac{0.3\mathcal{N}(x^{(0)} = 1; \mu = 1, \sigma_1 = 2)}{0.3\mathcal{N}(x^{(0)} = 1; \mu = 1, \sigma_1 = 2) + 0.7\mathcal{N}(x^{(0)} = 1; \mu = 1, \sigma_0 = 1)} \\
&= \frac{0.3\frac{1}{2\sqrt{2\pi}}e^{0}}{0.3\frac{1}{2\sqrt{2\pi}}e^{0} + 0.7\frac{1}{1\sqrt{2\pi}}e^{0}} \\
&= \frac{\frac{0.3}{2}}{\frac{0.3}{2} + 0.7} \\
&= \frac{0.3}{1.7} = \frac{3}{17} = 0.18
\end{aligned}
$$

$$
\begin{aligned}
r^{(1)} = P(z^{(1)} = 1|x^{(1)}) &= \frac{\theta\mathcal{N}(x^{(1)}; \mu, \sigma_1)}{\theta\mathcal{N}(x^{(1)}; \mu, \sigma_1) + (1-\theta)\mathcal{N}(x^{(1)}; \mu, \sigma_0)} \\
&= \frac{0.3\mathcal{N}(x^{(1)} = 3; \mu = 1, \sigma_1 = 2)}{0.3\mathcal{N}(x^{(1)} = 3; \mu = 1, \sigma_1 = 2) + 0.7\mathcal{N}(x^{(1)} = 3; \mu = 1, \sigma_0 = 1)} \\
&= \frac{0.3\frac{1}{2\sqrt{2\pi}}e^{-1/2}}{0.3\frac{1}{2\sqrt{2\pi}}e^{-1/2} + 0.7\frac{1}{1\sqrt{2\pi}}e^{-2}} \\
&= \frac{\frac{0.3\times 0.6}{2}}{\frac{0.3\times 0.6}{2} + (0.7 \times 0.1)} \\
&= \frac{0.18}{0.18 + 0.14} = \frac{18}{32} = 0.56
\end{aligned}
$$

Question 13 1 / 1 point

Suppose you are using a Gaussian Mixture Model (GMM) to model the density of a univariate (i.e., one-dimensional) dataset. The GMM has three components, and after training, you have obtained the following means and standard deviations, with each mixture component receiving equal weight in the mixture:

- Mixture Component 1:

$$\mu_1 = .5, \sigma_1 = 1$$

- Mixture Component 2:

$$\mu_2 = -.5, \sigma_2 = .2$$

- Mixture Component 3:

$$\mu_3 = 0, \sigma_3 = 1$$

Now, suppose you are given three new points

$$x^{(1)} = .4, x^{(2)} = -.5, x^{(3)} = 0$$

. Which of these points is assigned the highest likelihood by the GMM model? Which point is assigned the lowest likelihood?

○

$$x^{(1)}$$

has the highest likelihood and

$$x^{(3)}$$

has the lowest likelihood

○

$$x^{(3)}$$

has the highest likelihood and

$$x^{(1)}$$

has the lowest likelihood

○

$$x^{(3)}$$

has the highest likelihood and

$$x^{(2)}$$

has the lowest likelihood

✓ ○

$$x^{(2)}$$

has the highest likelihood and

$$x^{(1)}$$

has the lowest likelihood

▼ Hide Feedback

simply calculate the sum of contribution by three components at each of the given points to get

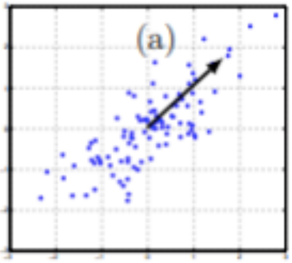
$$p(x^{(1)}) \approx 0.233, p(x^{(2)}) \approx 0.861$$

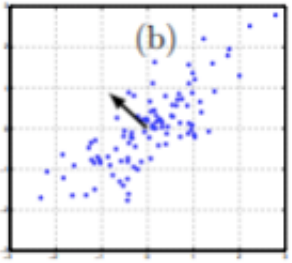
and

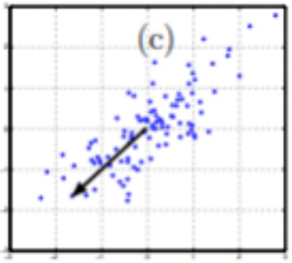
$$p(x^{(3)}) \approx 0.264$$

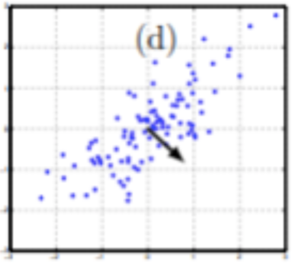
Question 14 1 / 1 point

Which of the following figures correspond to possible values that PCA may return for the **first** principle component?









○ (a)

○ (b)

○ (a), (b), (c) and (d)

✓○ (a) and (c)

Hide Feedback

The first principal component corresponds to the one with largest variance. In the options, only a and c have the largest variance.

Question 15 1 / 1 point

Suppose you are optimizing a linear regression function using the closed form approach on a dataset with D features (including the bias term) and N training examples. What is the time complexity of running leave-one-out cross validation on this training set?

○ $$O(D^3 + N + ND^2)$$

✓○ $$O(ND^3 + N^2D^2)$$

○ $$O(ND^2 + ND^3)$$

○ $$O(ND^2 + N^2D^3)$$

▼ Hide Feedback

this is the product of N (cost of cross validation) and the complexity of closed form solution (ND^2 + D^3)

Question 16 **1 / 1 point**

What is the decision boundary in the following model:

$$P(y = 1|x) = \frac{1}{1 + e^{-1-1.1x_1-5x_2}}$$

○ 1. The decision boundary is non-linear due to the non-linear form of the logistic function

○ 2. The decision boundary is linear and given by the equation

$$1 + 1.1x_1 + 5x_2 = .5$$

✓○ 3. The decision boundary is linear and it is given by

$$P(y = 1|x) = .5$$

○ None of the above

Question 17 **1 / 1 point**

You are running SGD with a large momentum of 0.9, and see the following gradients for a scalar parameter where {t} denotes the time-step:

$$\nabla J_{\mathbb{B}}(w^{\{0\}}) = 1$$

$$\nabla J_{\mathbb{B}}(w^{\{1\}}) = -1$$

$$\nabla J_{\mathbb{B}}(w^{\{2\}}) = 1$$

The update to the parameter at time 2 is given by

$$w^{\{3\}} = w^{\{2\}} - \alpha \Delta w^{\{2\}}$$

. What is the value of

$$\Delta w^{\{2\}}$$

? Start with

$$\Delta w^{\{0\}} = 1$$

and report your answer with two decimals (e.g., .02)

Answer: 0.82 ✗ **(.82)**

▼ Hide Feedback

$$\Delta w^{\{0\}} = 1$$

$$\Delta w^{\{1\}} = 0.9 * 1 + 0.1 * (-1) = 0.8$$

$$\Delta w^{\{2\}} = 0.9 * 0.8 + 0.1 * 1 = 0.82$$

Question 18 — 0 / 1 point

A function is convex if and only if

➡ ○ none of the above

○ it is monotonically increasing

✗ ○ its gradient always points towards the unique minimum

○ it has a single global minimum

Question 19 — 0.5 / 1 point

Which of the following is correct about bias-variance trade-off?

➡ ✓ □ Using a small k in k-NN increases the variance

✓ □ Pruning of a decision tree increases its variance

✗ □ For L2 loss the generalization error is the sum of our model's bias, variance and the irreducible error

➡ ✓ □ Validation set can be used to get an estimate for model's variance

▼ Hide Feedback

bias^2

Question 20 1 / 1 point

Given an NxD design matrix X, and labels y, which of the following techniques could potentially reduce the training error in logistic regression:

✓ □ Penalizing the model weights with L2 regularization

✓ □ Adding the feature 1 to each data point

✓ □ Select the best subset of features using L1 regularization and cross-validation

✓ □ Adding polynomial features to each datapoint

Attempt Score:	16.5 / 20 - 82.5 %
Overall Grade (highest attempt)**:**	16.5 / 20 - 82.5 %

Done