SYSC5405
Pattern Classification & Experiment Design

Final Exam Review

Prof James Green

jrgreen@sce.Carleton.ca

Systems and Computer Engineering, Carleton University

Format of the Final Exam

- Multiple choice quiz using BrightSpace
- 80 questions in 3 hours = ~2 min per question + 20 min for review
  - 5 questions per “page” = 16 pages
  - Can navigate freely through the 16 pages.
  - Order of questions randomized for each student
  - **Most** questions are multiple choice. A few calculated answers or mix/match.
- This is a closed book exam
  - You can bring a calculator, but will only need it for a couple questions
    - Calculator should be a simple calculator, no smartphones, etc.
  - No other resources may be used (no notes, no Internet access, etc.)
- Covers all topics in the course
  - Emphasis *roughly* matches in-class emphasis

1j) What kind of learning model would you use to teach a robot to play chess?

Supervised learning	**132979**
Semi-supervised learning	**132991**
Unsupervised learning	**133093**
Reinforcement learning	**133302**

1k) You train a machine learning model using all data that is available to you. You then test it on new data, collected over the following month and it works really well! This is because...

You tested the model on all the data originally available	**133318**
You used a GPU to achieve very deep learning	**133325**
The model generalized	**133969**
You got lucky!	**133993**

2e) If $f(x) = (x-5)^2 + 4$, then $5 =$

$argmax_x f(x)$	134013
$max_y f(x)$	134048
$argmin_x f(x)$	134303
$min_x f(x)$	134317

4b) Which of the following samples have the LOWEST entropy?

{rain, snow, snow, rain, snow}

{cat, no_cat, no_cat, no_cat, cat, cat, dog}

{pass, pass, pass, pass, pass, fail}

{a, e, b, b, a, d, c, c, d, b, e}

5d) For a K-NN classifier, decreasing K will:

Increase generalization of the classifier

Decrease computational complexity when classifying a new test point

Decrease training time

Decrease resilience to mislabelled training points.

5f) We are building a fish K-NN classifier that uses fish length (in inches) and weight (in grams). When we ship our system to a customer in the US, they re-train the system, but now measure weight in lbs (i.e. weight values are now much smaller).

The system is expected to no longer work well

The system is expected to work well only if mass was the most discriminating feature

The system is expected to work well only if they used Mahalanobis distance (scale by variance of data)

The system is expected to work well without any additional changes

6d) When optimizing a linear decision boundary in a six-dimensional feature space, the optimization criterion will have how many parameters?

12

6

13

7

6f) What is the definition of "learning rate"?

A parameter that controls the size of a parameter update in each epoch

A parameter that controls the amount of data to use in each epoch

The slope of the gradient in the current epoch

7c) Assume you have a categorical feature, colour, with values "red, green, blue, yellow, orange, purple". One-hot encoding would require:

One new feature

Three new features

Five new features

Seven new features

None of the above

7e) Assume you have the following feature data: $x = \{1, 1, 5, 9, 9\}$. Which of the following is the normalized version of x?

$x = \{1, 1, 5, 9, 9\}$ (i.e. x is already normalized)

$x = \{-0.5, -0.5, 0, 0.5, 0.5\}$

$x = \{0, 0, 0.5, 1, 1\}$

None of the above.

7f) Mahalanobis distance is equivalent to:

Euclidean distance, for 2 features or less

Euclidean distance after normalizing the data

Euclidean distance after standardizing the data

None of the above.

7g) Data imputation is useful when:

Feature data is missing during testing.

Feature data is missing during training.

We identify and censor outliers in the data.

All of the above.

None of the above.

8a) Which performance metric answers the question: "Out of all the positive predictions, what proportion are actually positive?"

a) Accuracy

b) Recall

c) Precision

d) MSE

8b) Which of the following performance metrics can be used for both regression and classification?

Accuracy

MSE

Precision

None of the above.

14a) The process of creating many new datasets, drawn randomly from a original training dataset, then training multiple weak classifiers, is called:

- Bagging
- Kernel regression
- Multi-label classification
- Maximum likelihood learning

14c) A classifier that easily generalizes to multi-class learning is:

K-nearest-neighbour

Support vector machines

Linear discriminant classification

None of the above

14d) You have been hired to classify incoming patients as to whether they will 1) be admitted or not, and 2) whether they will require an MRI. This is an example of:

A multi-class problem

A multi-label problem

A multi-objective problem

A multi-variate regression problem

16a) Active learning is a way to improve classifier accuracy:

through guided labeling of data

through resampling of the training data

by combining an ensemble of classifiers

without requiring additional data collection

16c) What is the purpose of the bottleneck layer in an autoencoder?

To allow the model to better memorize the inputs

To learn a concise representation of the input

To identify which training samples require labels

To add noise to the input to make the model more generalized

17a) Which of the following is NOT a method for combining models?

Averaging

Sampling

Voting

Stacking

17b) Which performance metric reflects the class imbalance of a test set?

Sensitivity

Precision

Specificity

AUC-ROC

17c) Undersampling is an approach to:

Create synthetic training data for the minority class

Early stopping during training

Create a balanced dataset

Training below computational limits of the processor

19a) Are Gaussian Mixture Models parametric or non-parametric?

Parametric

Non-parametric

Air Canada would like a system to predict how long flights will be delayed when passing through different Canadian airports. This is an example of which type of problem:

Classification	134350
Regression	134376
Reinforcement learning	134636
None of the above.	134708

Good luck on the exam!