COMP 350 **Applied Machine Learning**

Practice Final

December 03, 2018 *Fall 2018*

Instructions:

- If you believe a problem is incorrectly or incompletely specified, make a reasonable assumption and solve the problem. The assumption should not result in a trivial solution.
- In all cases, clearly state any assumptions you make in your answers.

Name	

Section	**Points Possible**	**Grade**
True/False	20	
Multiple Choice	20	
Debugging	20	
Coding	20	
Short Answer	20	
Total	100	

1 True/False [2 points each]

1.1. Given a trained word2vec CBOW model, it's easy to compute the vectors for out-of-vocabulary word. _____

1.2. In Latent Dirichlet Allocation, each document is assigned a single topic. _____

1.3. You can always extract as many principal components as there are input features. _____

1.4. Adding a batch normalization layer increases the number of parameters in a neural network. _____

1.5. ARI is a practical way of adjusting the number of clusters in K-Means for exploratory data analysis. _____

1.6. A Gaussian Mixture Model allows evaluating the probability of a new point under a fitted model. _____

1.7. The normalized mutual information (NMI) is not defined for cluster assignments with different numbers of clusters. _____

1.8. Isolation Forests assume Gaussian Distributed Data. _____

1.9. In a bag-of-word model with unigrams, using stop-words will reduce the number of features only marginally. _____

1.10. Convolutional layers in a NN typically have fewer parameters than densely connected layers. _____

2 Multiple Choice [5 points each]

Select all choices that apply.

2.1. Which of the following statements apply to neural networks?

(a) Fast to train on large datasets.

(b) Can learn arbitrarily complex functions.

(c) Work well when little training data is available.

(d) Provide state-of-the-art performance in computer vision and audio analysis.

(e) Have no hyper-parameters to tune.

2.2. Which of the following models requires solving an optimization problem (as opposed to a closed-form formula) to transform data?

(a) Non-Negative Matrix Factorization

(b) Latent Dirichlet Allocation

(c) PCA

(d) Linear Discriminant Analysis

(e) Paragraph Vectors

2.3. What are reasons to prefer Non-negative Matrix Factorization over PCA?

(a) Better reconstruction of the data.

(b) Sign of the components is meaningful.

(c) No cancellation effects.

(d) Can extract non-linear features.

(e) Faster.

(f) Deterministic results.

2.4. Which of the following cluster evaluation methods are unsupervised?

(a) Silhouette Score

(b) Adjusted Rand Index

(c) Normalized Mutual Information

(d) Stability based score

3 Debugging [10 points each]

For each code snippet, find and explain all errors in the task. Assume all necessary imports have been made. There can be more than one error per task!

(a) Task: Perform grid-search on a Keras sequential model for the number of units (50, 100, or 200) in the hidden layer. The network should be a one-hidden-layer network for 64 input features and 8 output classes. There are 3 bugs.

```
X_train, X_test, y_train, y_test = train_test_split(X, y)
model = Sequential([Dense(50),
Dense(8, activation="softmax"))
model.compile("adam", "multiclass_crossentropy", metrics=["accuracy"])
param_grid = {'hidden_units': [50, 100, 200]}
grid = GridSearchCV(model, param_grid)
grid.fit(X_train, y_train)
score = grid.score(X_test, y_test)
```

(b) Task: Write down the computation in a forward-pass of a feed-forward NN for classification with one hidden layer with 100 units, `tanh` non-linearity and a drop-out rate of 50% on the hidden layer. There are 2 bugs.

```
def forward(X, w1, b1, w2, b2):
  h1_net = np.dot(X, w1 + b1)
  dropout_mask = np.random.uniform(100) > .5
  h1_net[dropout_mask] = 0
  h1 = np.tanh(h1_net)
  out_net = np.dot(X, w2) + b2
  out_exp = np.exp(out_net)
  return out_exp - np.sum(out_exp)
```

4 Coding [10 points each]

Assume all necessary imports have been made.

(a) Define a multi-layer perceptron using the Keras Sequential interface with relu non-linearity and a single hidden layer with 100 hidden units for classifying the iris dataset.

(b) Apply PCA to detect outliers in a dataset given as X by reducing it to 10 dimensions. Assume there are 5% outliers. Include preprocessing.

5 Short Answer [5 points each]

(a) Explain the "CBOW" approach used in `word2vec`. How are the word representations found?

(b) Explain how "batch normalization" works.

(c) Compute the number of parameters in a convolutional NN with $16 \times 16 \times 1$ input, followed by two 3×3 convoluation layers with 4 maps each, followed by a 2×2 max pooling layer followed by an output layer with two units (don't forget biases). You can just write ou the multiplication and additions for each layer; you don't need to compute the additions and multiplications.

(d) Explain the generative process for a document in Latent Dirichlet Allocation.