Student Name _____ Student ID _____

Midterm Test _____ / 138

ECE521: Inference Algorithms and Machine Learning
Instructor: Prof. Brendan Frey
Friday, February 26, 2016, 3.10pm-5.00pm
Time: 110 minutes.
Total marks available: 76

Part I: True/False _____ / 42

For each question, circle one answer. If you select "Don't know", you will receive 2 marks, whereas if you select another answer and get it wrong, you will receive 0 marks, and if you get it right, you will receive 3 marks.

1. **True False Don't know** Using momentum with gradient descent helps to find solutions more quickly.

2. **True False Don't know** In neural networks, increasing the number of hidden units usually reduces the training error.

3. **True False Don't know** The L2 weight penalty is equivalent to a Laplace prior on the weights.

4. **True False Don't know** A classifier trained on less training data is less likely to over fit.

5. **True False Don't know** Convolutional neural networks have fewer parameters than fully connected networks because they use shared weights.

6. **True False Don't know** Stochastic gradient descent is better than batch gradient descent because it provides a more accurate estimate of the true gradient of the loss function.

7. **True False Don't know** It is best to keep the learning rate of a neural network constant as learning progresses.

8. **True False Don't know** A TensorFlow tensor with shape [1,2,3,4,5] can be added to a Tensor with shape [4,5] using broadcasting.

9. **True False Don't know** TensorFlow uses symbolic differentiation which results in speed-up in the execution of the code.

10. **True False Don't know** A TensorFlow session deploys the graph by binding it to a particular execution context (e.g. CPU, GPU).

Student Name _____ Student ID _____

11. **True False Don't know** In TensorFlow, it is illegal to add tensors from these two different graphs defined by tf.Graph().

12. **True False Don't know** In the TensorFlow graph, the edges represent the operations and the nodes represent the tensors.

13. **True False Don't know** In TensorFlow variable names can be scoped and the TensorBoard visualization uses this information to define a hierarchy on the nodes in the graph.

14. **True False Don't know** In TensorFlow, when computing the derivative of the loss function with respect to the weights, we do not have to explicitly pass the weights as an argument to the optimizer since TensorFlow automatically keeps track of the trainable variables.

Part I: Multiple Choice **_____ / 56**

For each question, circle one answer. If you select "I don't know", you will receive 4 marks, whereas if you select another answer and get it wrong, you will receive 0 marks, and if you get it right, you will receive 8 marks.

15. For which dataset is the nearest neighbor method most likely to *not* work well?

a) 1000 1-dimensional points, spread out uniformly in [0,1].

b) 1000 100-dimensional points spread out uniformly in $[0,1]^{100}$.

c) 1000 100-dimensional points spread out uniformly in a 2-dimensional subspace of $[0,1]^{100}$.

d) 100 2-dimensional points spread out uniformly in $[0,1]^2$.

e) I don't know.

16. Which of the following statements about the *k*-nearest neighbors method for a training set of size *N* is *true*?

a) As *k* grows from 1 to *N*, the classification accuracy of held out data consistently increases.

b) As *k* grows from 1 to *N*, the model overfits more.

c) The decision boundary is smoother for larger values of *k*.

d) If the data is not linearly separable, the method cannot achieve 100% training accuracy.

e) I don't know.

Student Name _____ Student ID _____

17. Which one of the following activation functions is *not* suitable for the backpropagation learning procedure?

a) 1/(1+exp(-z)).

b) tanh(z).

c) [z>0], where [True]=1 and [False]=0.

d) [z>0]z, where [True]=1 and [False]=0.

e) I don't know.

18. Consider a softmax model with a label k that can take on values 1, …, K:

$$P(k \mid x) = \exp(\sum_j w_{kj} x_j) \Big/ \sum_{k'=1}^{K} \exp(\sum_j w_{k'j} x_j) .$$

Which expression below is the derivative of the log-likelihood for a dataset $(x^{(1)}, y^{(1)}),...,(x^{(N)}, y^{(N)})$ wrt w_{kj} ?

a) $\sum_{i=1}^{N} \sum_{k=1}^{K} \left([y^{(i)} = k] - P(k \mid x^{(i)}) \right) P(k \mid x^{(i)}) \left(1 - P(k \mid x^{(i)}) \right) x_j^{(i)}$

b) $\sum_{i=1}^{N} \left(1 - P(y^{(i)} = k \mid x^{(i)}) \right) x_j^{(i)}$

c) $\sum_{i=1}^{N} \left([y^{(i)} = k] - P(k \mid x^{(i)}) \right) x_j^{(i)}$

d) $\sum_{i=1}^{N} \left(1 - P(y^{(i)} = k \mid x^{(i)}) \right) P(y^{(i)} = k \mid x^{(i)}) \left(1 - P(y^{(i)} = k \mid x^{(i)}) \right) x_j^{(i)}$

e) I don't know.

19. Which of the following datasets does not contain independent and identically drawn data?

a) 100 1-dimensional points sampled uniformly from the interval [0,5].

b) 100 2-dimensional points sampled uniformly from the circumference of a unit-radius circle.

c) 100 1-dimensional points with values 1, 2, 3, …, 99, 100.

d) 100 1000-dimensional points with the first dimension set to a value uniformly sampled from the interval [0,1] and each subsequent dimension set to the previous dimension plus Gaussian noise with std dev 0.01.

e) I don't know.

20. You would like to train a neural network with output $f(x)$ to predict a target y that has additive Laplacian noise u with pdf $P(u) = 0.5\ \lambda \exp(-\lambda|u|)$. Maximum likelihood training of this model is equivalent to minimizing which one of the following error functions?

a) $\sum_{i=1}^{N} \left| y^{(i)} - f(x^{(i)}) \right|$

b) $\sum_{i=1}^{N} \left(y^{(i)} - f(x^{(i)}) \right)^2$

c) $\sum_{i=1}^{N} u^{(i)} \left| y^{(i)} - x^{(i)} \right|$

d) $\sum_{i=1}^{N} \left| \lambda y^{(i)} - f(x^{(i)}) \right|^n$

e) I don't know.

21. Which one of the following procedures is a valid implementation of dropout?

a) During training, with probability 1-*keep_prob*, set the hidden activity to zero. At test time, multiply the outgoing weights by 1/*keep_prob*.

b) During training, with probability 1-*keep_prob*, set the hidden activity to zero. At test time, multiply the hidden activations by (*1-keep_prob*).

c) During training, with probability *keep_prob*, scale up each hidden activity by *1/keep_prob* and otherwise set it to zero. At test time, none of the activations or weights are scaled or dropped.

d) During training, with probability *keep_prob*, scale up each hidden activity by (*1-keep_prob*) and otherwise set it to zero. At test time, none of the activations or weights are scaled or dropped.

e) I don't know.

Part III: Short Answer **_____ / 40**

22. (8 marks) Write down the output produced by the following code:

```
from math import sqrt
nums = {int(sqrt(x)) for x in range(30)}
print nums
```

Answer_____

Student Name _____ Student ID _____

23. (8 marks) Write down the output produced by the following code:

```
a=np.array([1, 2, 3, 4, 5])
b=a.T
a[:]=0
print b
```

Answer_____

24. (12 marks) For logistic regression, show that the log-odds, log P(y=1|**x**)/P(y=0|**x**), is a simple linear weighted combination of the elements of the input vector **x**.

25. (12 marks) Design a neural network with one input, x, a single layer of two ReLU units, and one linear output unit, y, that perfectly fits the data (-5,1), (-4,1), (-3,1), (-2,1), (-1,2), (0,3), (1,4), (2,4), (3,4), (4,4), (5,4). *Enter the values of the weights in the boxes in the diagram below.*