CSCE883 Machine Learning Midterm **Name:**_____

Problem 1: (15 points)

(a) Explain why a too complex model such as very deep decision tree with limited samples will cause problems for classification performance

(b) 1) Describe what is overfitting and
2) plot how the training error and test error look like when overfitting happens;
3) describe how to detect overfitting during classifier training

(c) Briefly describe three ways of preventing overfitting in neural networks and why they work

Problem 2: (15 points)

Using Naïve Bayesian classifier to classify the following instance with 3 attributes (A1, A2, A3) into Mammal (M) or Non-mammal (N).

Instance: (A1=Yes, A2=No, A3=Yes).

Given: P(A1=yes|M)=0.2 P(A2=yes|M)=0.8, P(A3=yes|M)=0.4;
P(A1=Yes|N)=0.4 P(A2=yes|N)=0.4 P(A3=yes|N) =0.5
P(M)=0.3

Problem 3 (15 points)

Write down TP, FP, TN, FN into the corresponding cells of the below table and define the following criteria for evaluating classifiers

Accuracy (a) =

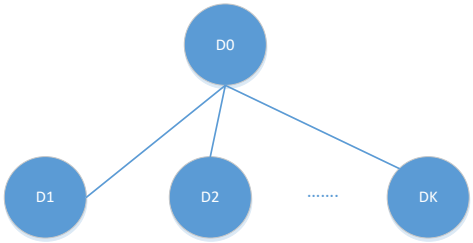
Precision (p)=

Recall (r) =

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes		
	Class=No		

Some papers tend to use the Accuracy as the only criterion for evaluating classifiers, describe the potential shortcoming of this criterion.

Problem 4 (10 points) Let p(i) denote the fraction of training instances belonging to a class i and K to be the number of classes . Write down the formulas for calculating the Entropy of this training set which is used in decision tree attribute selection.



Suppose the entropy of each node in the above decision tree is represented as $E(D_i)$ and the number of instance of each node is noted as $|D_i|$, write down the formula to calculate the information gain for the attribute used to split node D_0

Problem 5 (10 points):

The Pareto distribution has been used in economics as a model for a density function with a slowly decaying tail:

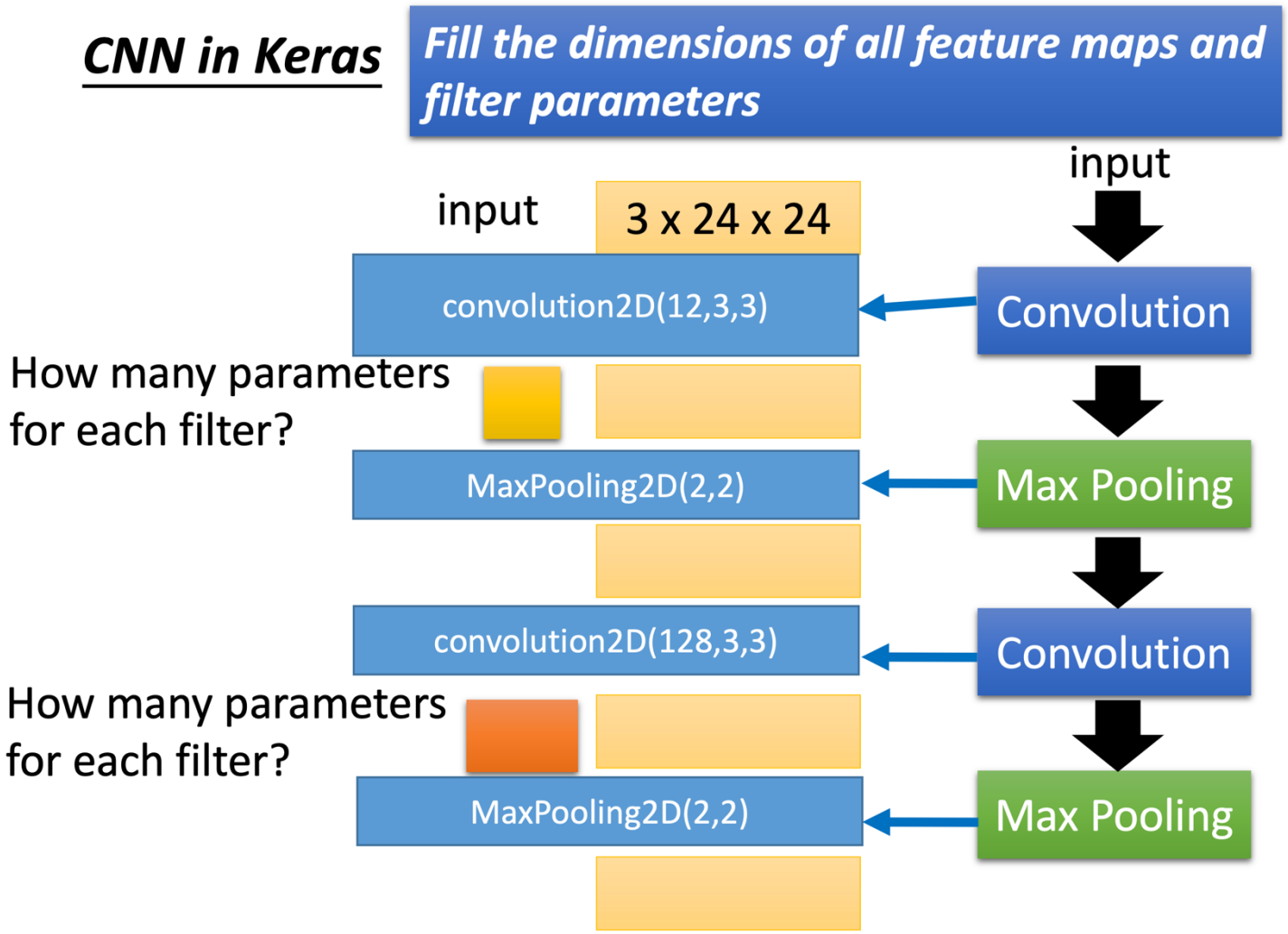
$$f(x|x_0,\theta) = \theta x_0^{\theta} x^{-\theta-1}, \quad x \geq x_0, \quad \theta > 1$$

Assume that $x_0 > 0$ is given and that $X_1, X_2, \cdots, X_n$ is an i.i.d. sample. Find the MLE of θ.

Problem 6 (35 points) Deep learning

a) Traditional multi-layer feedforward perceptron ANN has a notorious difficulty to train for models with multiple hidden layers. Describe how the recent deep learning algorithms solve this problem. (5points)

b) For the following CNN network, calculate the feature map dimensions and the parameter numbers for the filters. We use no padding and the stride is 1. (15 points)

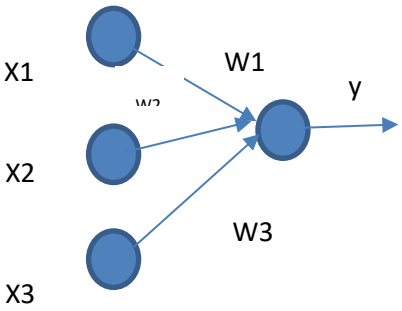


c) backpropagation (10 points)

Consider a 1-layer neural net with three input units, 1 output unit, no hidden units and no bias terms. Suppose that the output unit uses a sigmoid activation function, *i.e.*, $y = 1/(1+e^{-z})$, where z is the total input to the unit. Let y be the computed output of the neural net, let d be the desired output, and let $C = -d\log y - (1-d)\log(1-y)$ be the cross entropy error. Write down the equations for a single step of weight updates by gradient descent (based on a single data sample), and derive all the necessary derivatives. Simplify your answers, and be sure to clearly identify all the variables you use.

Hint: use the chain rule and recall the following results:

$$\frac{\partial y}{\partial z} = y(1-y) \qquad \frac{\partial \log u}{\partial u} = \frac{1}{u}$$



Only need to show $\frac{dC}{dw_1}$

d) What is the main function of a convolutional filter? What the main function of a max-pooling layer? (5 points)