Advanced Predictive Analytics

Mid Term Exam, Fall 2017

Total Marks = 40.

Time: 1 hr 15 mins.

Answer each problem in the space provided; use the back of the preceding sheet in extreme conditions only.
Declaration: By submitting this examination for grading, I affirm that I have neither given nor received help from another examinee.

Name: _____

Signature: _____

Q. 1: [10 pts total]

(a) (3+2 pts) (i) Suppose you are the using (multiple) linear regression model to solve a specific regression problem. Currently you are using 5 of the 10 available independent variables (predictors), and a training set of a given size. Indicate how would you expect the (i) model bias, (ii) model variance, and (iii) (model bias)^2 + model variance to be affected by changing the aspect mentioned in Col A, keeping everything else the same, by filling in the table below using:
I, D, DNC, ID for increase, decrease ,"does not change" and "it depends", as appropriate.
If you have stated ID for some entry, briefly justify your answer.

Col A	Bias	Variance	Bias^2 + variance
Increase training data size	DNC	D	D
Use 8 independent variables instead of 5.	D	I	ID

(ii) What is the impact of increasing the size of a "holdout" test dataset on the true (or "adjusted") R-square performance of your model?

No change. With more test data one can estimate the true error better, but the true error value itself does not change as the holdout data does not impact the developed model in any way.

(b) (3pts). Briefly, what do you understand by the statement that the multilayered perceptron (MLP) with a single hidden layer is a universal approximator?

MLP with a single hidden layer can uniformly approximate any continuous function on a compact input domain to arbitrary accuracy provided the network has a sufficiently large number of hidden units and appropriate parameters. This result holds for a wide range of hidden unit activation functions.

(c) (2 pts) State two properties of Naïve Bayes that make it so suitable for parallel/distributed computation and for analyzing "streaming" data (e.g. each document is processed only once and cannot be further stored or revisited).

Each P(x_j|C_i) can be computed independently.
Each computation is the ratio of counts, and counts are easy to parallelize and/or implement in a distributed fashion, and moreover, can trivially be updated incrementally (for streaming data)

Q 2. (10 pts)

(a) (3pts). Mathematically show that the classification boundary returned by logistic regression for a 2-class problem is linear in the space of the independent variables.

At boundary P(y|x)=constant (0.5 for Bayes decision, some other constant if costs are asymmetric),
so log-odds = sum(xi*wi) = constant, i.e. linear boundary perpendicular to weight vector.

(b) (3 pts). Mention two situations where you may prefer to use stochastic gradient descent (SGD) to determine the parameters of a multiple linear regression model instead of solving the (batch) least squares problem?

Ans:

Any two of the following will suffice.
(i)Large dataset. Batch too expensive
(ii)Non-stationary problem, relationship between x and y changes with time, so need to adapt online.
(iii)Data is not available all at once but is streaming

(c) (2 pts) Mention two distinct uses of K-fold cross-validation.

a) Used to set hyperparameters such as the amount of regularization to be used or number of training epochs for an MLP.
b) Used to estimate the true performance, ie expected error on future data.

(d) (2pts) Briefly explain how the slack penalty "C" can be used to control the bias-variance tradeoff in SVMs.

The SVM cost function minimizes 0.5*w_i^2 + C*slack_penalty.
The slack penalty is the cost of a data sample being on the wrong side of the +/- decision boundary. A higher C will heavily penalize misprediction and hence mould the decision boundary to closely fit the data. This will lead to an overfit model (high variance and low bias). A lower C means the objective will depend on the w_i term more than the slack term. This can lead to underfitting. (low bias and high variance). Thus C acts as an inverse regularization penalty

The tradeoff can be visualized as follows. A low C will create a rigid boundary with a uniform margin resulting in many points being misclassified. As C increases, misclassification is penalised more and the decision boundary becomes more flexible by reducing the margin.

Q 3. (10 pts)

(a) (3 pts). How will you formulate a 4-class classification problem so that you can solve it using logistic regression? You'll need to use 4 - 1 = 3 models. Choose any of the 4 classes to be the base class C_k. Set all coefficients for C_k to 0 to make the system identifiable. Then for the other classes (indexed by i), calculate ln(P(C_i | X) / P(C_k | X)). Finally for each datapoint, assign it the class that has the highest value from the three models with respect to C_k. If all models given negative results, C_k is the class to be assigned.

.

(**b). (2 pts).** Give two situations (with reasoning) where you may prefer to use Support Vector Regression instead of Multiple Linear Regression.

- When data has outliers, SVR is more robust to them and hence preferred.
- When the data being fitted is highly nonlinear, SVR is prefered since using the kernel trick, it can find a linear boundary in a suitable higher-dimensional space without explicitly mapping the features into this space. This allows it to obtain non-linear boundaries in the original feature space

(c) (2+3pts) Consider the Bayesian belief network below with the associated conditional probability tables (CPT)s.
(i) Why is the Cloudy variable not present in the CPT for "WetGrass"?
(ii) What is the probability that the sprinkler was ON given that it was a cloudy day and it rained and the grass is wet?

i) Cloudy is not present because Wet Glass is conditionally independent of Cloudy given its parents Sprinkler and Rain.

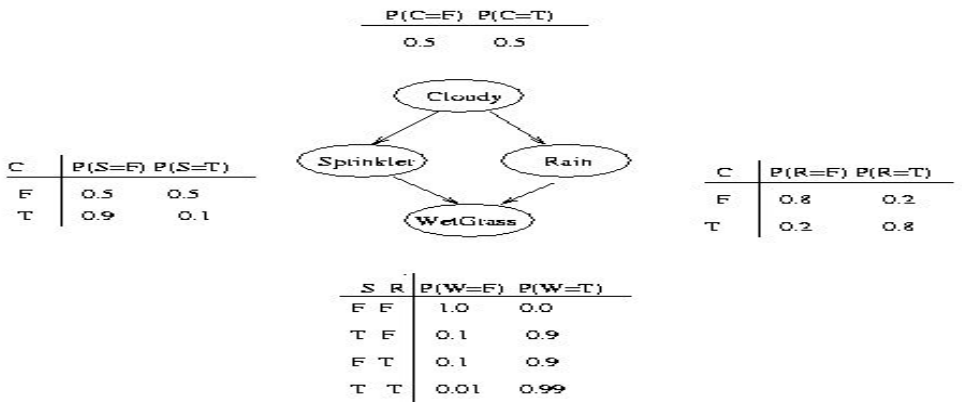
ii) $$P(S=T \mid C=T, R=T, W=T) = \frac{P(S=T, C=T, R=T, W=T)}{P(C=T, R=T, W=T)}$$

$$= \frac{P(S=T, C=T, R=T, W=T)}{P(S=T, C=T, R=T, W=T) + P(S=F, C=T, R=T, W=T)} = \frac{.0396}{.0396 + .324} = 0.1089$$

where

$P(S=T, C=T, R=T, W=T) = P(S=T \mid C=T) * P(C=T) * P(R=T \mid C=T) * P(W=T \mid S=T, R=T)$
$= .1 * .5 * .8 * .99 = 0.0396$

$P(S=F, C=T, R=T, W=T) = P(S=F \mid C=T) * P(C=T) * P(R=T \mid C=T) * P(W=T \mid S=F, R=T)$
$= .9 * .5 * .8 * .9 = .324$



Q3 (10 pts total)

(a) (1.5x4=6 pts) Choose the best alternative. If multiple answers are correct, select all the correct ones. 1.5 for correct answer, -0.5 for incorrect answer; 0 for no answer.

(a) The LMS algorithm (also called "delta rule") for ADALINE
   a. Is based on SGD
   b. Provides one way of updating the weights of a linear regression model in an online fashion
   c. Uses momentum to accelerate convergence
   d. Is guaranteed to converge to the optimal solution after which no more weight updates will occur.

(b) The (two-class) LDA (Linear Discriminant Analysis) classifier
   a. Assumes that the decision boundary is linear, and so is ideal if the classes are actually linearly separable
   b. Is a generalization of the QDA classifier
   c. Estimates an input covariance matrix using all the training data
   d. Provides Bayes optimal decisions if adequate training data is available

(c) In SVMs, the "kernel trick" allows one to
   a. obtain non-linear boundaries in the original feature space
   b. maximize the margin between two classes
   c. minimize the number of support vectors
   d. find a linear boundary in a suitable higher-dimensional space without explicitly mapping the features into this space.

(d) In an ADABOOST based ensemble of N classifiers
   a. The classifiers can be learnt in parallel on different boostrapped samples of the training data
   b. Can reduce both bias and variance
   c. Uses decision stumps as the base classifiers
   d. Training is done through gradient boosting in function space

(b) (2+2=4 pts).

(i) What is the key difference between Random Forests and Bagging?

(ii) Suggest one dataset (in terms of some relevant property) for which Random Forest is likely to perform worse than bagging if you have a limit (say 20) to the maximum number of classifiers that can be used in the ensemble.

i) Random Forests pick from a random subset of the features at each split for each bootstrap sample whereas bagging examines all the features and then picks which features to split on. This additional randomization has the goal of decreasing correlation among trees without affecting bias too much

ii) A dataset where only a few of all the features provide information to predict the dependent variable. In this case, random forests may grow trees with splits based on features that don't provide enough information on the class and thus you may end up with weaker learners than the ones averaged from bagging alone.