CS446: Machine Learning (SOLUTIONS) **Fall 2017**

October 12^{th}, 2017

- This is a closed book exam. Everything you need in order to solve the problems is supplied in the body of this exam.
- This exam booklet contains **four** problems. You need to solve all problems to get 100%.
- Please check that the exam booklet contains **15** pages.
- You have 75 minutes to earn a total of 100 points.
- Answer each question in the space provided. If you need more room, write on the reverse side of the paper and indicate that you have done so.
- **Besides having the correct answer, being concise and clear is very important. For full credit, you must show your work and explain your answers.**

Good Luck!

Name (NetID): (1 Point)

Logistic Regression		/25
Naive Bayes		/25
Decision Trees		/25
Short Answer		/24
Total		/100

1 Logistic Regression (25 pts)

4 Free Points!

(a) (1 pts) Briefly describe the difference between the generative approach and the discriminative approach for probabilistic classifiers.

The Discriminative approach models $P(\mathbf{Y}|\mathbf{X})$ directly whereas the generative approach models $P(\mathbf{Y},\mathbf{X})$ and then conditions on X to derive $P(\mathbf{Y}|\mathbf{X})$. (Optional: Naive bayes is an example of a generative model whereas Logistic regression is an example of a discriminative model.)

Rubric:
Define what discriminative and generative approach models (1pt)

(b) (3 pts) Choose from the table below and fill in the blank.

Logistic Regression is an example of a ______**discriminative**______ model that models $P(y|\mathbf{x},\mathbf{w})$ according to a ______**Bernoulli**______ distribution. It is best suited for a ______**binary classification**______ problem.

discriminative	estimation	Bernoulli
generative	Poisson	creative
binary classification	regression	normal

Rubric:
Each correct answer gets (1pt)

(c) (2 pts) Why is the sigmoid function used in logistic regression? Give two reasons. Hint: Consider representation and optimization.

The sigmoid funciton maps $\mathbb{R} \rightarrow [0,1]$, which is a good representation of probability. It is also smooth and differentiable, which is preferred for optimization.
Rubric:
Each valid reason gets (1pt)

(d) (15 pts) For the following questions, assume we have a dataset of size n with $y_i \in \{0,1\}$ and $\mathbf{x}_i \in \mathbb{R}^d$. Also assume we are modeling $P(y_i = 1|\mathbf{x}_i, \mathbf{w}) = \mathbf{sigm}(\mathbf{w}^\intercal \mathbf{x}_i)$ for $1 \leq i \leq n$.

(i) (2 pts) Write down the logistic regression decision rule for maximizing accuracy.

$$\hat{y}(\mathbf{x}_i) = \begin{cases} 1 & \mathbf{sigm}(\mathbf{w}^\intercal \mathbf{x}_i) > \theta \\ 0 & otherwise \end{cases}$$

Where θ is the decision boundary. Any $0 < \theta < 1$ is correct.
Rubric:
Correct decision labels (1pt)
Correct conditions (1pt)

(ii) (6 pts) Write down the conditional log likelihood of Y conditioned on $\mathbf{X}$ and $\mathbf{w}$, $LL(\mathbf{w})$, for logistic regression in terms of $\mathbf{sigm}(\mathbf{w}^\intercal \mathbf{x}_i)$, y_i, and n.

$$\begin{aligned} L(\mathbf{w}) &= p(Y|\mathbf{X}, \mathbf{w}) \\ &= \prod_{j=1}^{n} p(y_j|\mathbf{x}_j, \mathbf{w}_j) \\ &= \prod_{j=1}^{n} p(y_j = 1|\mathbf{x}_j, \mathbf{w})^{y_j} p(y_j = 0|\mathbf{x}_j, \mathbf{w})^{1-y_j} \\ &= \prod_{j=1}^{n} (\mathbf{sigm}(\mathbf{w}^\intercal \mathbf{x}_j)^{y_j} (1 - \mathbf{sigm}(\mathbf{w}^\intercal \mathbf{x}_j))^{1-y_j} \\ LL(\mathbf{w}) &= \sum_{j=1}^{n} y_j log(\mathbf{sigm}(\mathbf{w}^\intercal \mathbf{x}_j)) + (1 - y_j) log(1 - \mathbf{sigm}(\mathbf{w}^\intercal \mathbf{x}_j)) \end{aligned}$$

Rubric:
Correct L(w) (2pts)
Correct LL(w) (4pts)

(iii) (7 pts) In addition to the assumption of $P(y_i = 1|\mathbf{x}_i, \mathbf{w}) = \mathbf{sigm}(\mathbf{w}^\intercal \mathbf{x}_i)$, we now want to regularize our $\mathbf{w}$ by imposing a gaussian prior, $p(\mathbf{w}) = \mathcal{N}(0, \lambda^{-1}I) = \sqrt{\frac{\lambda}{2\pi}} e^{-\frac{\mathbf{w}^\intercal \mathbf{w}\lambda}{2}}$, where I is the identity matrix and λ is a positive number.

- (5 pts) Suppose we are interested in the MAP estimate of $\mathbf{w}$, derive the new log likelihood, $LL'(\mathbf{w})$.

$$L'(\mathbf{w}) = p(Y|\mathbf{X}, \mathbf{w})p(\mathbf{w})$$

$$= \prod_{j=1}^{n} (\mathbf{sigm}(\mathbf{w}^\intercal \mathbf{x}_j)^{y_j} (1 - \mathbf{sigm}(\mathbf{w}^\intercal \mathbf{x}_j))^{1-y_j} \sqrt{\frac{\lambda}{2\pi}} e^{-\frac{\mathbf{w}^\intercal \mathbf{w}\lambda}{2}}$$

$$LL'(\mathbf{w}) = \sum_{j=1}^{n} y_j log(\mathbf{sigm}(\mathbf{w}^\intercal \mathbf{x}_j)) + (1 - y_j) log(1 - \mathbf{sigm}(\mathbf{w}^\intercal \mathbf{x}_j)) - \frac{\lambda \mathbf{w}^\intercal \mathbf{w}}{2} + const$$

Rubric:
Error carried forward from previous part(ECF) (-1pt)
MAP Estimate (3pts)
Correct LL'(w) (2pt)

- (2 pts) Derive the gradient descent/ascent update rule for our new regularized logistic regression in terms of $\nabla LL(\mathbf{w}'_k), \eta, \lambda, \mathbf{w}'_k$.

Let

$$LL'(\mathbf{w}) = LL(\mathbf{w}) - \frac{\lambda \mathbf{w}^\intercal \mathbf{w}}{2} + const$$

be the new objective function. Then,

$$\nabla LL'(\mathbf{w}) = \nabla LL(\mathbf{w}) - \lambda \mathbf{w}$$

And the gradient ascent update rule is then

$$\mathbf{w}'_{k+1} = \mathbf{w}'_k + \eta(\nabla LL(\mathbf{w}'_k) - \lambda \mathbf{w}'_k)$$

Rubric:
Correct derivation of $\nabla LL'(w)$ (1pt)
Correct derivation of the gradient descent/ascent update rule (1pt)

2 Naive Bayes (25 pts)

(a) (2pt) Is the following statement correct? Please write True or False, and briefly explain your answer. Incorrect explanation will result in 0 points.

A Naive Bayes classifier is trained on a dataset which satisfies the naive Bayes assumption, i.e. conditional independence. As the training dataset's size grows to infinity, the trained model will eventually always achieve zero TRAINING error.

False: If the true classes are not perfectly separable, then there is always a non-zero probability of error due to label noise (distributions join at tails). Other valid reasons will get full points as well as long as they makes senses.

(b) (2pt) Is the following statement correct? Please write True or False, and briefly explain your answer. Incorrect explanation will result in 0 points.

A Naive Bayes classifier is trained on a dataset which satisfies the naive Bayes assumption, i.e. conditional independence. As the training dataset's size grows to infinity, the trained model will eventually always achieve zero TESTING error, given that the naive Bayes assumption is true on testing data as well.

False: same reason as above.
Other valid reasons will get full points as well as long as they makes senses.

(c) (5pt) You are asked to pick a ball from one of three colored jars, one of which is green. The probability of selecting a ball from each jar is uniform. If the overall probability of picking a red ball is $\frac{3}{4}$, the probability of a red ball from the green jar is $\frac{1}{2}$. What is the probability that you selected a green jar given the pick of a red ball? Either show the exact computation or briefly explain how you derived the result.

$$P(G \mid R) = \frac{P(R, G)}{P(R)} = \frac{P(R \mid G)P(G)}{P(R)} = \frac{\frac{1}{2}\frac{1}{3}}{\frac{3}{4}} = \frac{2}{9}$$

For partial credits:
Stating $P(R \mid G) = \frac{1}{2}$ will get 1 point.
Stating $P(G) = \frac{1}{3}$ will get 1 point.
Correctly stating Bayes formula and apply will get 2 point.
Correct answer will get another 1 point.

(d) In this problem, we will train a probabilistic model to classify a CS graduate S is from UIUC or UMich, based on their grades X on 4 common courses.

For each student s_i, he/she is either from I for UIUC or M for UMich, with grades $\mathbf{x}_i = \{x_1, x_2, x_3, x_4\}$, where x_i denote their grades on course C_i, and all the grades follows the Bernoulli distributions, supposing there are only 2 grades, A or B. Given S_i,

$$P(x_i = A \mid s_i = I) = \alpha_{iI}$$

$$P(x_i = A \mid s_i = M) = \alpha_{iM}$$

$$P(x_i = B \mid s_i = I) = \beta_{iI}$$

$$P(x_i = B \mid s_i = M) = \beta_{iM}$$

We want to build a naive bayes classifier to compute $P(s_i = I \mid \mathbf{x})$.

(i) (3 pts): Using naive Bayes assumption, write down an expression for $P(s_i = I \mid \mathbf{x})$ using the following terms $P(s_i = I), P(s_i = M), P(x_i \mid s_i = I), P(x_i \mid s_i = M)$

$$P(s_i = I \mid \mathbf{x}) = \frac{P(s_i = I) \prod_{i=1}^{4} P(x_i \mid s_i = I)}{P(s_i = I) \prod_{i=1}^{4} P(x_i \mid s_i = I) + P(s_i = M) \prod_{i=1}^{4} P(x_i \mid s_i = M)}$$

For partial credits:
Correct Bayes formula on x including the expanded denominator will get 1 point.
Correct independent assumption expansion on the joint probability (nominator) will get 1 point.
Overall correctness get another 1 point.

(ii) (5 pts): How many parameters MUST be estimated to train such a naive Bayes classifier? Please list all of them using probability expressions or/and α, β. Note: not all parameters are required to be estimated for the predictive model. (2 pts for correct number, 3 pts for correct listing.)

There are total 9 parameters needs to be estimated: $P(s = I), \alpha_{1I}, \alpha_{2I}, \alpha_{3I}, \alpha_{4I}, \alpha_{1M}, \alpha_{2M}, \alpha_{3M}, \alpha_{4M}$.
For partial credits:
α_{iI} **and** β_{iI} **are not listed together, or stating they are duplicated (similarly** α_{iM}**) will get 1 point.**
Listing $P(s = I)$ **or** $P(s = M)$ **will get 1 point.**
Overall correctness get another 1 point.

(iii) (5 pts) Suppose you observe 4 students, 3 from UIUC and 1 from UMich. Among 3 UIUC students, 2 of them have A grades over all 4 courses, and 1 of them has B grades over all 4 courses. The UMich student has B grades over all 4 courses. Suppose this is all the data you have, using the MLE, what's the probability that all B grade student comes from UMich? Either show the exact computation or briefly explain how you derived the result.

From data, for UIUC student, each course has $\frac{1}{3}$ chance to get B, and for UMich Student, each course has 1 chance to get B. $P(M \mid B) = \dfrac{P(B \mid M)P(M)}{P(B)}$

$$\frac{(1/4)}{(1/4)+(3/4)*(1/3)^4} = \frac{27}{28}.$$

For partial credits:

Stating $P(B \mid M) = 1$ will get 1 point.

Stating $P(M) = \frac{1}{4}$ will get 1 point.

Correct naive Bayes formula on x including the expanded denominator will get 1 point.

Overall correctness get another 2 point.

(iv) (3 pts) Someone argues that the Naive Bayes assumption is not appropriate for this question. Do you agree with this? Explain your answer.

Not appropriate. Students get 3 As will have higher possibility to get another A in real life. The grades of courses for each student are not conditionally independent given the their school.

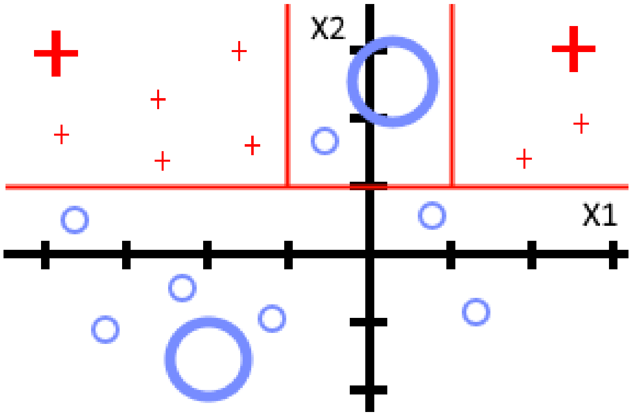
Appropriate. (can still get full credit as long as explanation makes senses). An example is, one may argue that given school is fixed, the difficulty of each course can be assumed fixed (have similar statistic in each semester). Since each course covers a different field of knowledge, grades in different courses can be assumed conditionally independent.

For partial credits:

The following answer "It's not appropriate because the naive Bayes assumption assumes conditional independence, and this is not true." will receive 0 credits because it only restates the question with the "condition independence" information, which has been already provided at the part(a) as well. We expect you to explain more on why you think this assumption is not true or this assumption is true.

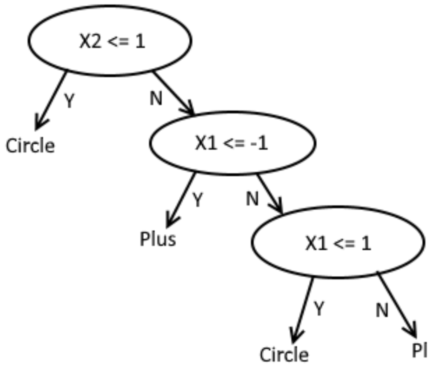
3 Decision Trees (25 pts)

(a) (8 pts) You are given a dataset where each sample has two continuous-valued features, x_1 and x_2, and a label, either a $+$ or a $\circ$. The dataset has been plotted for you below. Each tick on the axis is one unit.



(i) (6 pts) Using tests in the form of either $\{x_1 \leq a\}$, $\{x_1 \geq a\}$, $\{x_2 \leq a\}$, or $\{x_2 \geq a\}$, where $a \in \mathbb{Z}$, draw the decision tree that can correctly classify all 14 points in the figure using as few tests as possible. Be sure to **clearly label** which branch of the test corresponds to Yes and No. (Trees with fewer tests will earn more credit.)

One acceptable answer is given below. There are others, but the best trees use three tests.



(ii) (2 pts) Draw each region in the decision tree on the plot above. For each region, write the majority label for points in that region. Points were assigned on a variety of factors, but in general, 1 point was deducted for each additional test included and 1 point was deducted for each training example that was misclassified.

See above. Note that your plot may vary based on your answer to the part (i). This part was graded based with respect to your answer to part (i). In the optimal case, you should have four regions.

(b) (8 pts) Consider three models trained over the same dataset. Model A is a single decision tree (with no depth limit or pruning), model B is trained using the Adaboost algorithm with decision stumps, and model C is a random forest. Compare the bias and variance of these models by filling in the blanks below with either $<$, $\approx$, or $>$, and provide a brief justification of your choice. If you would need more information to answer the problem, write "?" and state your reasoning. Each comparison is worth 1 point and each justification is worth 1 point.

(i) $Bias(B)$ ____$\approx$____ $Bias(A)$

Justification: **A will have a low bias since it is not pruned. The individual stumps in B will have a high bias but the overall model will have a low bias. Therefore, $Bias(A) \approx Bias(B)$.**

(ii) $Var(B)$ ____$<$____ $Var(A)$

Justification: **The variance of a decision tree with no depth limit or pruning is very high. However, each decision stump has low variance and weighting their votes together will only decrease the variance.**

(iii) $Bias(C)$ ____$\approx$____ $Bias(A)$

Justification: **A will have a low bias since it is not pruned. Each individual tree has a low bias and bagging does not increase the bias, so we can say that C will have a low bias as well.**

(iv) $Var(C)$ ____$<$____ $Var(A)$

Justification: **Although the variance of each individual tree may be large, since we are using many of them, we would expect the variance to be lower with an ensemble method.**

(c) (9 pts) Boosting with Decision Trees

(i) (3 pts) When we use boosting with decision trees, we train a series of shallow decision trees (decision stumps) that, when combined, form a strong learner. In order for us to say this, what condition must the error of each of these shallow trees satisfy?

We need the Weak Learning assumption, that is, that the error rate of each individual each weak learner is slightly better than chance, i.e., $Error[L_W] = Error[L_{random}] - \epsilon$**. Full points were awarded if it was indicated that the error of each weak learner is less than (but not equal to)** 0.5**.**

(ii) (2 pts) When training Adaboost, we want to **decrease or leave** (increase /decrease/leave) the weights of the samples that our weak learner predicted correctly, and we want to **increase** (increase/decrease/leave) the weights of the samples that our weak learner predicted incorrectly.

(iii) (4 pts) On the next page, we have provided you with a skeleton for the algorithm for binary Adaboost. Refer to the table below. For each blank line, write the letter corresponding to the appropriate expression.

(a) $\log \frac{error_i}{1-error_i}$	(e) $\exp(\alpha_i \mathbb{1}_{\mathbf{x} \text{ correctly predicted by } tree_i})$
(b) $\log \frac{1-error_i}{error_i}$	(f) $w_{\mathbf{x}} \exp(\alpha_i error_i)$
(c) $\exp(\frac{error_i}{1-error_i})$	(g) $w_{\mathbf{x}} \exp(\alpha_i \mathbb{1}_{\mathbf{x} \text{ correctly predicted by } tree_i})$
(d) $\exp(\frac{1-error_i}{error_i})$	(h) $w_{\mathbf{x}} \exp(\alpha_i \mathbb{1}_{\mathbf{x} \text{ incorrectly predicted by } tree_i})$

Algorithm 1 Binary Adaboost Training

```
n ← number of training examples
k ← number of stumps to train
max_depth ← the depth of each stump
for $\mathbf{x} \in$ data do
    $\mathbf{w_x} \leftarrow \frac{1}{n}$
end for
for $i \in \{1, \cdots, k\}$ do
    $tree_i \leftarrow$ train_stump(data, max_depth)
    $error_i \leftarrow$ error(tree, data)
    $\alpha_i \leftarrow$ (1) ____________ (b) ____________
    for $\mathbf{x} \in$ data do
        $w_{\mathbf{x}} \leftarrow$ (2) ____________ (h) ____________
    end for
    for $\mathbf{x} \in$ data do
        $w_{\mathbf{x}} \leftarrow w_{\mathbf{x}} / \sum_{\mathbf{x}' \in \text{data}} w_{\mathbf{x}'}$
    end for
end for
Return $(tree_i, alpha_i)$ for $i \in \{1, \cdots, k\}$
```

4 Short Answer Questions (24 pts)

Some useful guidelines for this section.

- If the question is asking to select multiple choices, you will get points only when all the choices are correct.
- If the question is asking whether a statement is true or false, then you will get points only when you explain your choice. Writing only true or false will get you 0 pt.

(a) (2 pts) **(Select One)**

What are the priors on the weights which correspond to L_1 and L_2 regularization?

(i) Bernoulli for L_1 and Uniform for L_2.

(ii) Uniform for L_1 and Bernoulli for L_2.

(iii) Gaussian for L_1 and Laplace for L_2.

(iv) Laplace for L_1 and Gaussian for L_2.

iv

(b) (2 pts) **(Select One)**

Suppose you figured out that the labels of nearby points in your training data are very different. Which of the following options would you consider in k-Nearest Neighbor model to improve the performance?

(i) Increase the value of k.

(ii) Decrease the value of k.

(iii) The performance of the model with noisy data does not depend on k.

(iv) None of the above.

(i)

(c) (2 pts) **(Select One)**

Suppose you want to learn a logistic regression model using gradient descent. You run gradient descent for 100 iterations with the learning rate, $\alpha = 0.1$, and compute the negative of the log-likelihood after each iteration. You observe that the negative of the log-likelihood decreases quickly and then levels off after some iterations. What can you conclude from this observation?

(i) A larger value for the learning rate α is required.

(ii) A smaller value for the learning rate α is required.

(iii) $\alpha = 0.1$ is an effective choice of the learning rate.

(iv) None of the above.

(iii)

(d) (3 pts) **(Select Multiple)**

Suppose the data is being generated from some probability distribution and you created two Decision Tree models; DT_1 and DT_2. DT_1 is trained using finite number of training examples, whereas DT_2 is trained using infinite number of training examples. In comparison to DT_1, DT_2 will have:

(i) lower variance.

(ii) same variance.

(iii) lower bias.

(iv) same bias.

(i) and (iv)

(e) (3 pts) **(Select Multiple)**

Which of the following is / are ensemble technique(s)?

(i) Random Forest.

(ii) Singular Vector Machine.

(iii) Adaboost.

(iv) Neural Network.

(i) and (iii)

(f) (3 pts) Write two ways to avoid overfitting?

Any two of the followings:

(a) Go for simpler models.

(b) Reduce variance by taking into account fewer variables and parameters.

(c) Use cross-validation.

(d) Use regularization techniques (such as L1 penalty, L2 penalty, etc.)

1.5 points are given for each correct answer.

(g) (3 pts) When we minimize the sum of squared errors using gradient descent in the Linear Regression problem, we can get multiple local optimum solutions. Is this statement true or false? (You can assume that the number of data points is larger than the number of features.)

False. The objective function in Linear Regression is strongly convex. (Full points are given even if the student has written only convex)

(h) (3 pts) Suppose you and three of your friends are working on the CS446 project, individually. You developed a new learning model which helped you to achieve the best predictions for the given dataset. In order to beat your model predictions, your friends came up with their own models and claimed that they are better than yours. With whom would you agree and why?

Friend A: "Your model is nowhere near to my model. Look at the training error rates in my model! They are well below than yours."

Friend B: "Your model is just nothing in front of mine! Look at the test error rates! I got these results for the best value of α, chosen by experimenting on the test data."

Friend C: "My model is better than yours. Look at the test error rates! I got these results for the best value of α, chosen with 5-fold cross validation."

I would agree with C. A is claiming better error rates on training data which is not a useful metric to compare. B claims to have better error rates on the test data, but she has used test data as validation set. So, it is essentially not capturing generalization error.
1 point for agreeing with C.
1 point for writing what is wrong with A (or writing why C is better on the same arguments).
1 point for writing what is wrong with B (or writing why C is better on the same arguments).

(i) (3 pts) Your friend shared a fantastic experience working on CS446 project and then asked you a question. She said, "In order to reduce overfitting, I restricted my Decision Tree model to have a small maximum depth. It seems to work much better on the test data, but I really don't know why; what do you think?". What answer would you give in order to clarify her confusion?

3 points: Having a bound on the depth reduces variance, hence it is generalizing well.
3 points: Having a bound on the depth reduces complexity of the model, hence it is generalizing well.
1 point: Deeper trees are more prone to noise in the data.
1 point: Shallow trees generalize better than deeper trees.
0 point: It reduces overfitting. (That is already given in the question. One needs to justify why it is working better on test data.)

This page is intentionally left blank. You can use it for scratch paper.