COMS W4701: Artificial Intelligence, Spring 2022

Homework 6

Instructions: Compile all solutions to the written problems on this assignment in a single PDF file. **Show your work by writing down relevant equations or expressions, and/or by explaining any logic that you are using to bypass known equations**. For coding problems, we recommend that you write, label, and comment all of your code in one Jupyter notebook file. Follow the submission instructions to submit all files to Gradescope. Please be mindful of the deadline, as late submissions are not accepted, as well as our course policies on academic honesty.

You will be working with the following dataset. There are three trinary features x_1, x_2, and x_3, and two classes 0 and 1.

Sample	x_1	x_2	x_3	y
1	$+1$	-1	-1	0
2	$+1$	0	$+1$	0
3	0	0	-1	0
4	0	-1	$+1$	0
5	$+1$	$+1$	$+1$	1
6	0	$+1$	0	1
7	0	0	$+1$	1
8	-1	0	$+1$	1

Problem 1: Naive Bayes (25 points)

(a) Write a small program "training" a naive Bayes model using the provided data. You can simply hardcode the parameters since they can be computed by inspection, but we recommend that you store them neatly and also allow for smoothing to be added in a later part. Which parameters are 0 with no smoothing ($\alpha = 0$)?

(b) Identify **all** combinations of feature inputs for which our model will predict zero likelihood for both classes using the parameters learned above.

(c) Compute the distribution $\Pr(Y|x_1, x_2, x_3)$ for each training sample. Show a plot of the eight distributions (e.g., sample on x-axis, $\Pr(Y = 0)$ and $\Pr(Y = 1)$ as two separate lines on y-axis).

(d) Retrain the naive Bayes parameters using Laplace smoothing with $\alpha = 1$. Repeat part (c) with the new smoothed parameters. Briefly compare and contrast your observations about the distributions with and without smoothing.

(e) For either model, on which two training samples would we feel the most uncertain about our predictions? Briefly explain, referencing the specific feature combinations.

Problem 2: Decision Tree (20 points)

Suppose we train a decision tree model using the provided training data.

(a) Show all information gain computations that would be considered when learning the model. You do not have to show all of the numerical calculations, but please show the expressions that would be used. Sketch the resulting decision tree (ties may be broken arbitrarily).

(b) Now consider training a "random" forest of three decision trees, where each tree is trained on all data using the three different subsets of two features, i.e., (x_1, x_2) only, (x_1, x_3) only, and (x_2, x_3) only. Sketch each of the learned trees (no need to show calculations). What is the training accuracy of each tree individually? What is the highest attainable training accuracy of the forest if we classify using a majority vote?

Problem 3: Linear Classifier (25 points)

Suppose we train a linear classifier with a hard threshold activation function using the perceptron algorithm on the provided training data. We predict $y = 0$ for $h_{\mathbf{w}} < 0$ and $y = 1$ for $h_{\mathbf{w}} > 0$. We will look at the implication of the tiebreaker decision being assigned to each class.

(a) Plot the data as points in x_1-x_2-x_3 space, using different colors and markers to indicate the two different classes. Are the data linearly separable? You may have to try different viewpoints to get a full perspective of the space.

(b) Write a small program implementing perceptron, starting with weights $\mathbf{w} = (0, 0, 0, 0)$ and using learning rate $\alpha = 1$. Learn two models, one in which the tiebreaker $h_{\mathbf{w}} = 0$ predicts $y = 0$, and one in which it predicts $y = 1$. Track the weights over each iteration until convergence. Generate two plots, one for each model, of the four weight components (you can plot them as four separate lines in one graph). Remember that convergence is achieved only when the classifier is correct on a full pass through the training data.

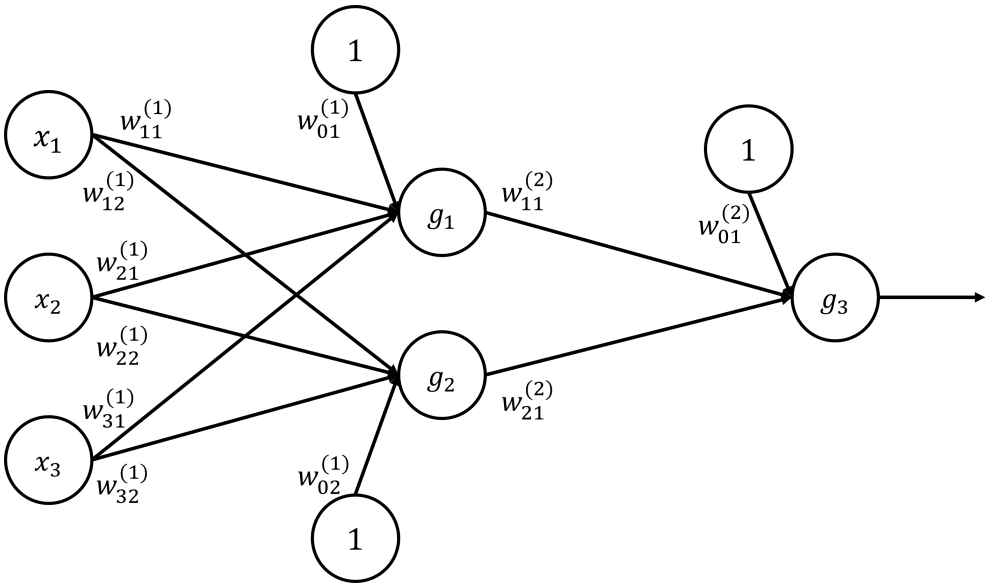
(c) Experiment with increasing and decreasing the learning rate α for each model. How does α affect the final learned weights and convergence speed?

(d) Compute and plot the sigmoid values of each of the training data using the two learned models (i.e., sample on x-axis and value on y-axis, one plot for each model). Which data predictions do we feel the most uncertain about, and what are their associated probabilities?

Problem 4: Neural Network (30 points)

Consider training a simple neural network model for classification of the training data set. We will be using the network shown below, with a three-unit input layer, two-unit hidden layer, and one-unit output layer. We will initially refer to the activation functions generically as g_1, g_2, and g_3. Each takes in a bias component in addition to the outputs from the previous layer.

Unlike the models in the previous problems, we will simply use the output of g_3 as the output of our classifier, which we can interpret as a probability value instead of a class value.



(a) In terms of the inputs, weights, and activation functions, write expressions for each of the two hidden layer outputs as well as the output of the output layer.

(b) We use the squared loss function $L(h_{\mathbf{w}}) = \frac{1}{2}(y - \hat{y})^2$. Write expressions for $\frac{\partial L}{\partial w}$, for each of $w = w_{01}^{(1)}$, $w = w_{11}^{(1)}$, $w = w_{01}^{(2)}$, $w = w_{11}^{(2)}$, in terms of the network parameters and outputs (do not leave derivatives in your expressions).

(c) Write a program implementing backpropagation to train the entire network, starting with all weights equal to 0. Use $\alpha = 1$ and the standard sigmoid for each activation function. Iterate through each of the training data, and compute and plot the loss over each iteration until convergence. You should select a suitable threshold at which to stop.

(d) Experiment with increasing and decreasing the learning rate α (try at least a factor of 2). Show two loss plots, one result from using $\alpha > 1$ and one result from using $\alpha < 1$. How does α affect the final learned weights and convergence speed?

(e) Experiment with using different activation functions: tanh, softplus, and ReLU. If you find that your losses are not converging, you may have to decrease the learning rate to a suitable amount. Show a convergent loss plot for each and briefly describe your observations.

Submission

You should have one PDF document containing all solutions, responses, and plots. You should also have a Python file or notebook containing all code that you wrote for each of the problems. Submit the document and code file to the respective assignment bins on Gradescope. For full credit, you must tag your pages for each given problem on the former.