1. (35pts) **Pre-midterm**

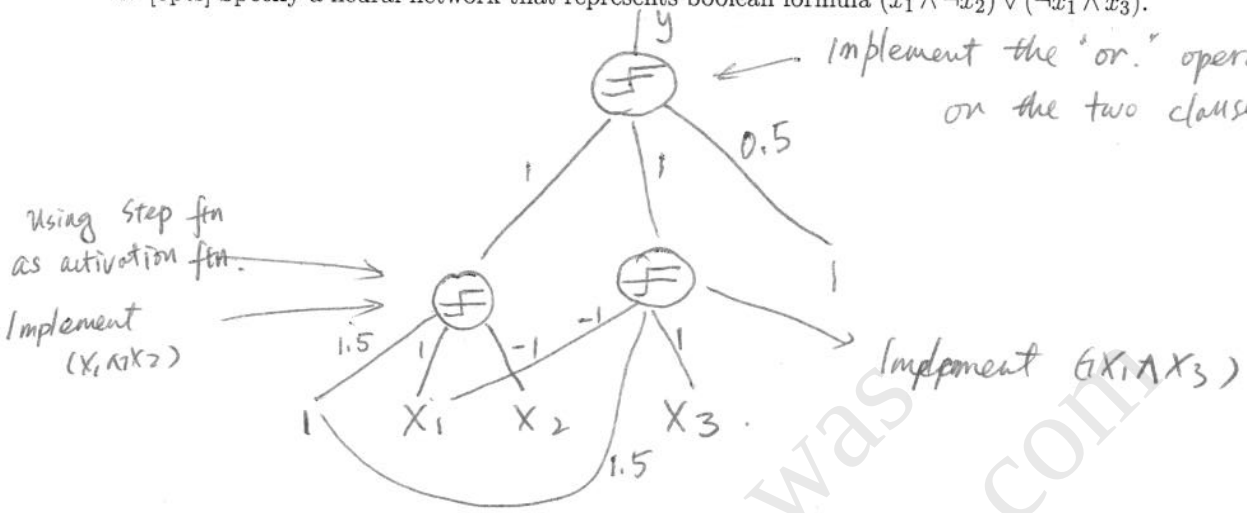
   a. (4pts) For a neural network, which of the following choices most strongly affects the trade-off between under-fitting and over-fitting:

      i. The initial weights
      ii. The learning rate
      iii. The number of hidden nodes
      iv. The choice of the online or batch learning algorithm

   b. (6pts) When using generative models for classification, we assume the data is generated by a procedure defined by the generative model. Describe the generative procedure for text data when using a Naive Bayes classifier (using the bernoulli distribution).

First generate class label based on p(y).
Given y, for each word in vocabulary, decide if it is included in the document according to the bernoulli distribution of that word for that class.

e. [6pts] Specify a neural network that represents boolean formula $(x_1 \wedge \neg x_2) \vee (\neg x_1 \wedge x_3)$.



f. (8pts) **True or False**:

i. Larger training sets lead to higher probability of overfitting.

False.

ii. For the K-nearest neighbor algorithm, larger K values will lead to higher probability of overfitting.

False.

iii. Logistic regression learns a non-linear decision boundary because the logistic function is non-linear.

False.

iv. Given linearly separable data, online Perceptron only finds a local optimal (not globally optimal) solution of the hinge loss function, because different initializations of the weights can lead to different solutions.

False.

b. (4pts) Which of the following algorithms do you expect to see most performance improvement when used with bagging? Briefly explain you answer in terms of stability of the algorithms.

i. Decision stump

ii. Decision tree with pruning

iii. Decision tree without pruning

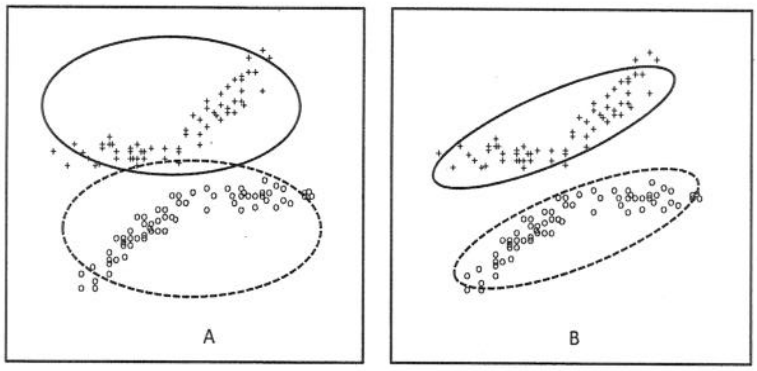
bagging reduces variance, works best with low bias, high variance (less stable) algorithms.

c. (4pts) If the training set contains noise in class labels, which ensemble learning method do you expect to be hurt more by the label noise, boosting or bagging? Why?

Boosting. It can increase the weights of the noise examples significantly.

3. (8pts)(**Gaussian Mixture Models**) In the following figures, '+' and 'o' denote positive and negative examples respectively. Three students fit Gaussian Mixture Models to this data.

a. Students A and B fit one Gaussian to each class. The solid (dashed) ellipse indicates the probability contour for the positive (negative) class. Which model do you prefer and why? What constraints do they each use for the covariance matrix?

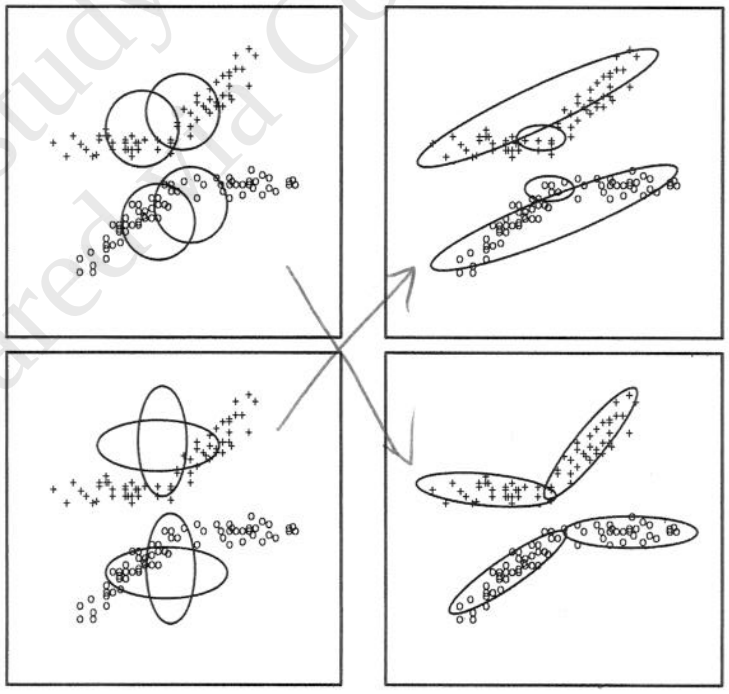


B is better fit.

For A, the covariance matrix is constrained to be diagonal and identical.

For B, only identical.

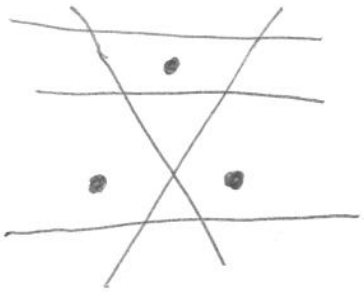
b. Student C fit two Gaussian to each class and used EM to learn the parameters. The left column below shows two initializations for EM and the right column shows the models after the 1st iteration. Match each initialization with its successive model.



4. (18pts) **Learning theory**

a. [5pts] Let H_{2d} be the hypothesis space of all possible 2-dimensional linear threshold units. Prove that $\text{VC}(H_{2d}) \geq 3$.

To prove $VC(H_{2d}) \geq 3$, we only need to show a set of 3 pts that can be shattered by H_{2d}.



This shows ~~that~~ all possible bi-partitions of the three pts, each achieved by a hypothesis in H_{2d}.

Thus H_{2d} shatters these three points.

b. [4pts] Suppose that you are able to find a set of 4 instances that cannot be shattered by H_{2d}. Does this imply that $\text{VC}(H_{2d}) < 4$? Explain.

No. To show $VC(H_{2d}) < 4$. one has to prove there exists no 4 points that could be shattered by H_{2d}.