# DS-GA-1003: Machine Learning (Spring 2021)

Final Exam (May 13 – May 14)

- You should finish the exam within **2 hours** once it is started and submit on Gradescope by **5:00pm EST on May 14**.

- You can refer to textbooks, lecture slides, and notes. However, searching answers online and collaboration are not allowed.

- You can write your solution either directly on the this sheet or on a separate sheet.

- Please leave enough time to upload your solution to Gradescope. You can submit it as a PDF file or any image format accepted by Gradescope. Please make sure the writings are legible.

- We suggest you save a timestamp of your start time, so that if you failed to submit in time (we hope not!), you can send us the answer sheet by email with the start and end timestamps.

Question	Points	Score
Bayesian methods	14	
Multiclass	8	
Trees	12	
Boosting	9	
Neural networks	10	
Latent variable models	13	
Total:	66	

1. **Bayesian methods.** Your are given the following dataset of N pairs of scalar real values $\mathcal{D} = \{x_i, y_i\}_{i=1}^{N}$:
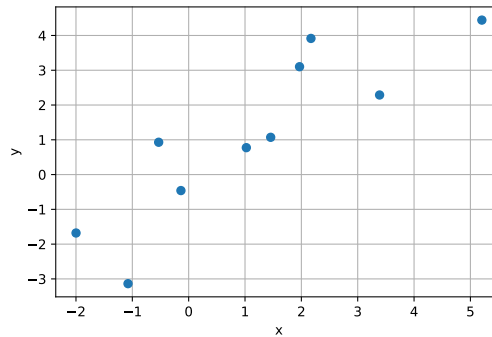
You would like to learn to predict y given x. Since you have little data you think it would be best to fit a simple model, yet data are not perfectly regular, so you settle for a Gaussian probabilistic conditional model with a linear predictor:

$$p(y|x, w) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-wx)^2} \text{ parametrized by } w \in \mathbb{R}.$$

Before seeing any data, an expert had told you that in this problem it can basically go two ways, in half of the cases y is typically equal to x or y is typically equal to $-x$. You interpret this information by using a binary prior on w:

$$p(w) = \begin{cases} 1/2 & \text{if } w = 1 \text{ or } w = -1 \\ 0 & \text{otherwise} \end{cases}.$$

(a) (2 points) Give the mathematical expression of the prior predictive function $y \mapsto p(y|x)$ describing your belief on y given a value of x taking only the probabilitic model into account (and not the data $\mathcal{D}$).

(b) (2 points) Assuming $x = 2$, sketch the graph of the prior predictive function $y \mapsto p(y|x)$ for $y \in [-10, 10]$ (you should note that it is a Gaussian mixture).

(c) (2 points) You now consider the data $\mathcal{D} = \{x_i, y_i\}_{i=1}^N$. Give the mathematical expression of the likelihood of the entire dataset $L(w, \mathcal{D})$ as a function of the likelihood per sample $p(y_i|w, x_i)$. Which assumption are you implicitly making?

(d) (4 points) Recall the definition of the posterior $p(w|\mathcal{D})$ as a function of $L(w, \mathcal{D})$ and $p(w)$ up to a multiplicative constant that is independent of w. Give the expression of $p(w|\mathcal{D})$ for the model considered here (Gaussian probabilistic conditional model and binary prior).

(e) (2 points) Knowing that $\sum_{i=1}^{N}(y_i + x_i)^2 = 229.4$ and $\sum_{i=1}^{N}(y_i - x_i)^2 = 12.9$, what is $\hat{w}_{\mathrm{MAP}}$, the maximum a posteriori estimate of w, and why?

(f) (2 points) Assume now that we are fixing w to $\hat{w}_{\text{MAP}}$. We draw a new value of x. We want to reduce the predictive function $y \mapsto p(y|x, \hat{w}_{\text{MAP}})$ to a point estimate, meaning that we want to use a single value $\hat{y}$ as prediction instead of the full distribution. How would you proceed and what would be the value of $\hat{y}$ you would choose as a function of $\hat{w}_{\text{MAP}}$ and x?

2. **Multiclass loss function.** Consider a multiclass classification problem with input $x \in \mathcal{X}$ and output $y \in \mathcal{Y}$ where $\mathcal{Y} = \{1, \ldots, K\}$. Let $\varphi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$ be the feature map. Alice proposed to reduce the multiclass classification problem to binary classification in the following way. First, map each example (x, y) to $K - 1$ examples: for each $k \in \mathcal{Y} \setminus \{y\}$, create a new binary classification example where the input is

$$\varphi(x, y) - \varphi(x, k)$$

and the output is $z_k = +1$.

(a) (3 points) Suppose we use hinge loss for the binary classification problem. Write down the loss function $\ell(x, y, w)$ of a single example (x, y) where $w \in \mathbb{R}^d$ denotes the weight vector. Note that you need to first map (x, y) to binary examples then sum the hinge loss of each binary example.

(b) (2 points) Given the binary classifier w, how would you predict the multiclass label for a test example x? Write an expression for the prediction $\hat{y}$.

(c) (3 points) Recall the generalized hinge loss for multiclass classification:

$$\ell_{\text{hinge}}(x, y, w) = \max_{y' \in \mathcal{Y}} \mathbb{I}\left[y \neq y'\right] - w \cdot (\varphi(x, y) - \varphi(x, y')),$$

where $\mathbb{I}(\cdot)$ is the indicator function. Let the groundtruth label be $y = 3$ and the total number of classes be $K = 4$. In which of the following cases would $\ell(x, y, w) = \ell_{\text{hinge}}(x, y, w)$? **Select all that apply.**

	$w \cdot \varphi(x, 3)$	$w \cdot \varphi(x, 1)$	$w \cdot \varphi(x, 2)$	$w \cdot \varphi(x, 4)$
A	3.25	2.82	1.20	1.74
B	3.25	2.82	3.11	4.60
C	3.25	2.10	0.81	1.93

3. **Decision trees.** Consider the following dataset with input $x \in \mathbb{R}^3$ and $y \in \{0, 1\}$. Both x_i and y are binary random variables.

x_1	x_2	x_3	y
0	0	0	0
0	0	1	1
0	1	0	1
1	1	1	0

You will use *Gini index* as the impurity measure to construct a decision tree.

(a) Compute the weighted average of node impurities when splitting on each of the three features. Note that for each feature, since the value is binary, there is only one way to split. Your final answer should be a real number.

    i. (2 points) Split on x_1

ii. (2 points) Split on x_2

iii. (2 points) Split on x_3

(b) (3 points) Draw the full decision tree and write the prediction for each leaf node. You can split ties arbitrarily.

(c) (3 points) Is the above decision tree contructed using the greedy approach optimal (i.e the minimum depth tree with zero training error)? If yes, explain in one sentence; if not, draw a decision tree with smaller depth and zero training error.

4. **Adaboost.** Consider the following binary classification dataset:
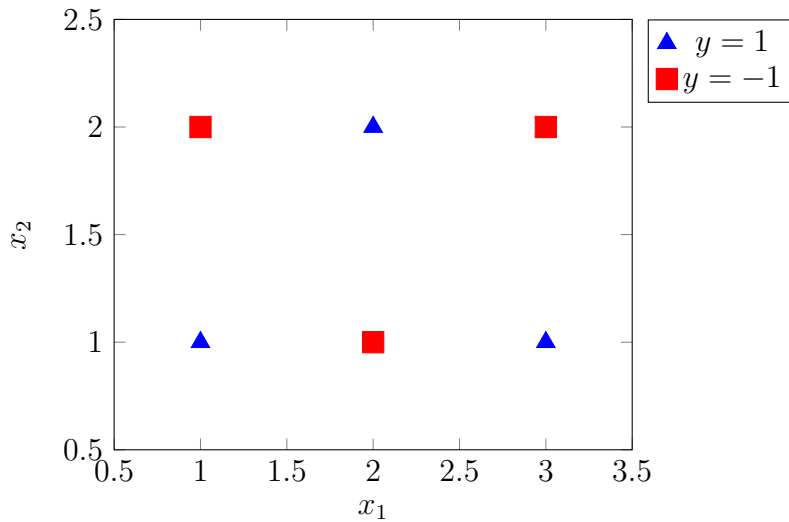

Alice plans to use Adaboost to solve the problem but she is not sure which base/weak classifier to use. For each of the base classifier choice below, (1) explain if Adaboost (using the base classifier) can reach 100% accuracy on the dataset; (2) if yes, draw the decision boundary in the first iteration and circle the points whose weights will increase in the next iteration.
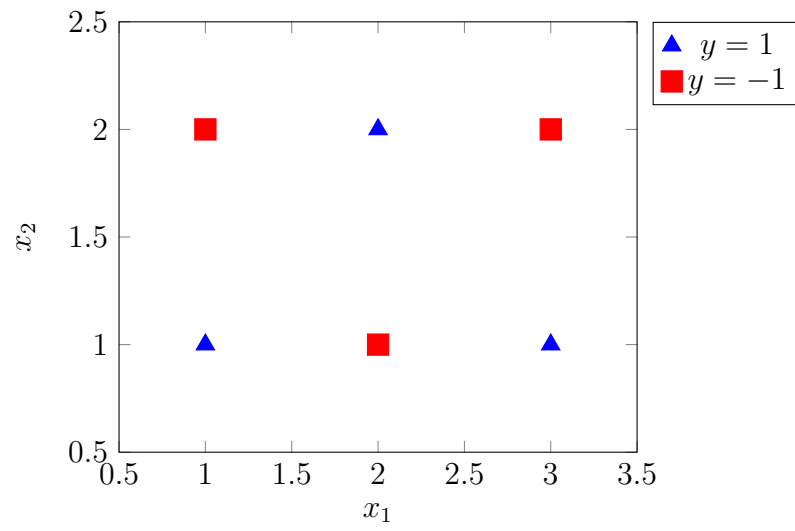
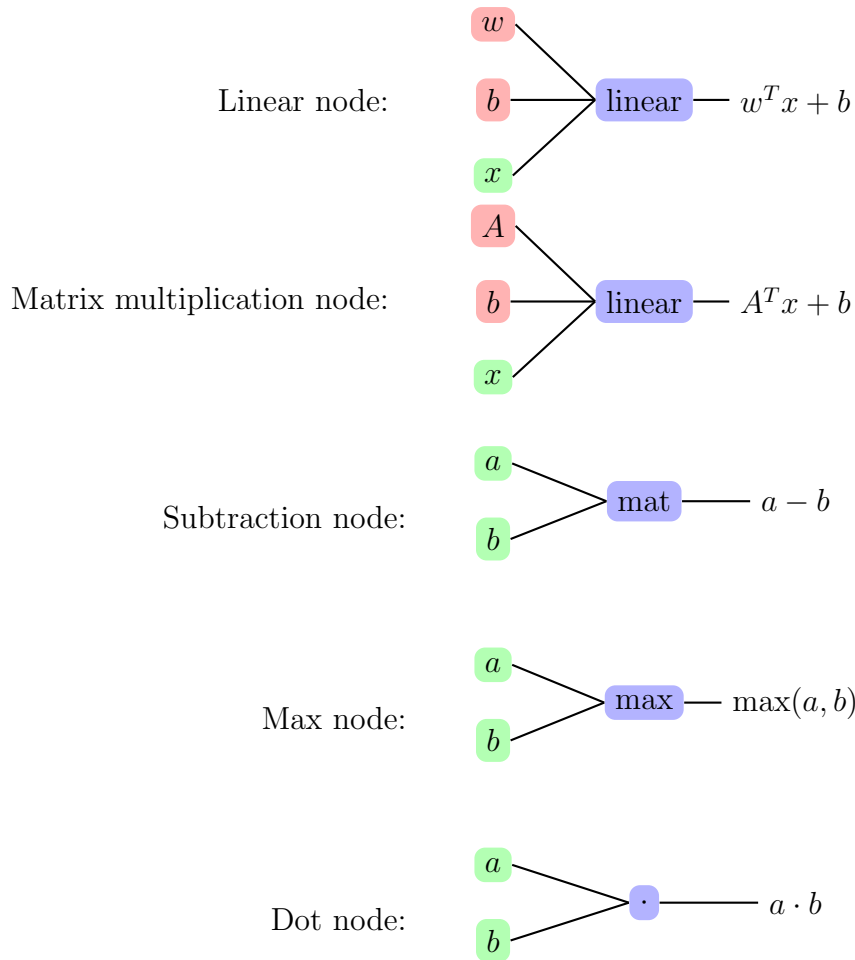(a) (3 points) Linear classifier: $f(x) = w_1 x_1 + w_2 x_2 + b$
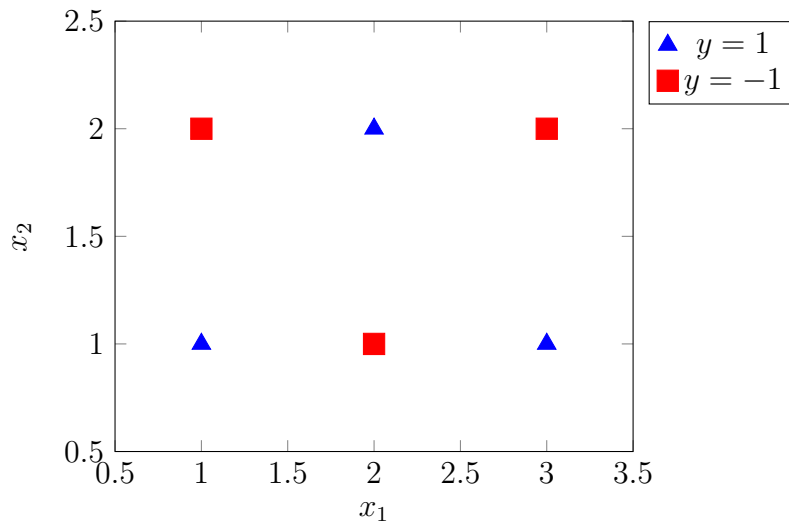
(b) (3 points) Decision stump: depth-1 decision tree
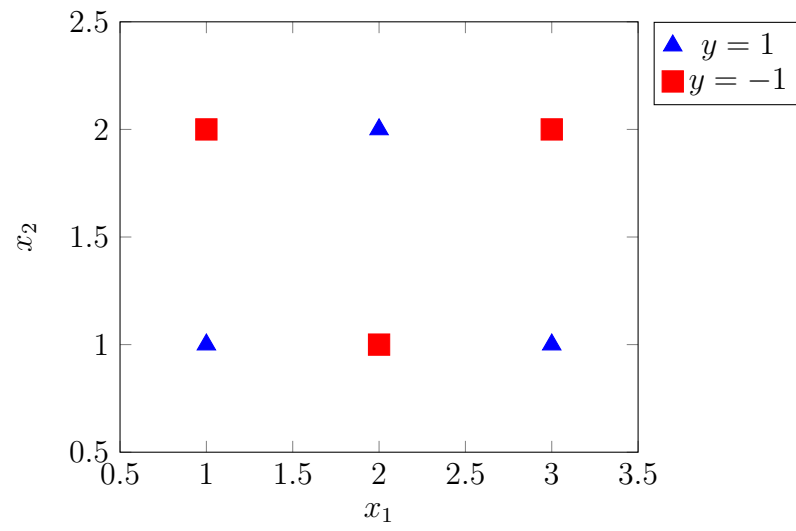
(c) (3 points) Quadratic classifier: $f(x) = w_1 x_1^2 + w_2 x_2^2 + w_3 x_1 x_2 + b$

5. **Neural networks.** Consider the following types of nodes in a computation graph. Blue, red, and green nodes represents functions, parameters, and inputs respectively. You do *not* need to color the nodes when drawing computation graphs though.

Linear node:

w
b → linear — $w^T x + b$
x

Matrix multiplication node:

A
b → linear — $A^T x + b$
x

Subtraction node:

a
b → mat — $a - b$

Max node:

a
b → max — $\max(a, b)$

Dot node:

a
b → $\cdot$ — $a \cdot b$

Note that if the input are vectors, the max node outputs the maximum value of the two vectors for each coordinate.

(a) (2 points) Consider the following computation graph:
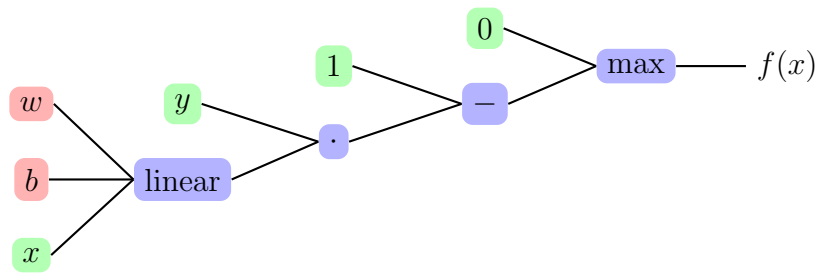
Give an expression for $f(x, y)$.

(b) (1 point) Let x and y denote the input and label for a binary classification task. What is the name of the loss function given by the function f above?

(c) (4 points) Modify the above computation graph to compute the hinge loss of a two-layer (one hidden layer) neural network with ReLU activation, and circle the node corresponding to the activation function. Please only use the computation nodes given above.

(d) (3 points) Give an expression for the output of your computation graph in the previous question.

6. **Gaussian mixture models.** Consider a Gaussian mixture model with two mixture components:

$$z \sim \text{categorical}(\pi_1, \pi_2),$$
$$x \mid z = i \sim \mathcal{N}(\mu_i, \sigma_i).$$

where $\mu_i \in \mathbb{R}$. Suppose we know the true value of the following parameters:

$$\sigma_1 = \sigma_2 = 1,$$
$$\pi_1 = \pi_2 = 0.5.$$

Now we want to estimate μ_i given some observed data using the EM algorithm.

(a) Suppose at initialization, we have $\hat{\mu}_1 = \hat{\mu}_2 = 0$.

    i. (2 points) What is the posterior of z estimated in the E-step? Give expressions for $p(z = 1 \mid x)$ and $p(z = 2 \mid x)$. Your final answer should be a real number.

ii. (3 points) What is the updated value of $\hat{\mu}_1$ and $\hat{\mu}_2$ in the M-step? You can assume that we have infinite data and your expression can contain μ_1 and μ_2.

iii. (2 points) After the update, the marginal likelihood of the data will (You can assume that the updated values are different from the initialization)
    A. increase
    B. decrease
    C. increase or decrease

iv. (2 points) What is the value of $\hat{\mu}_1$ and $\hat{\mu}_2$ at convergence (i.e. the marginal likelihood no longer changes)?

(b) (2 points) Given the above result, suggest a better initialization strategy.

(c) (2 points) Recall that in the EM algorithm, we need to decide the number of components beforehand. Suppose we try to estimate *three* components given data generated by a Gaussian mixture of *two* components. The marginal likelihood is more likely to

      A. increase

      B. decrease

Congratulations! You have reached the end of the exam.