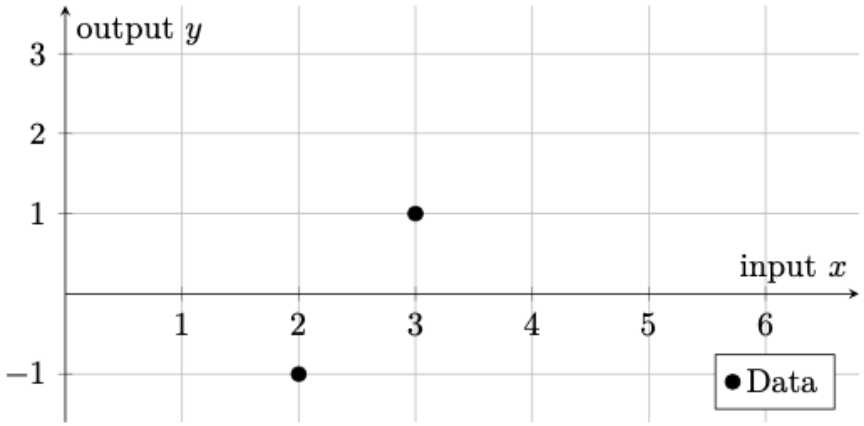
GHENT UNIVERSITY

ARTIFICIAL INTELLIGENCE (E016350)
Prof. Aleksandra Pizurica
AY 2024/2025

Intro to supervised learning – Part 1

1. **Linear regression**

   (a) Assume that you record a scalar input x and a scalar output y. First, you record $x^{(1)} = 2$, $y^{(1)} = -1$, and thereafter $x^{(2)} = 3, y^{(2)} = 1$. Fit a linear regression model $\hat{y} = h_{\mathbf{w}}(x) = w_0 + w_1 x$ using the squared error loss. Use the model to predict the output for the test input $x_t = 4$, and add the model to the plot below:



   (b) Now, assume you have made a third observation $y^{(3)} = 2$ for $x^{(3)} = 4$ (is that what you predicted in (a)?). Update the parameters $\mathbf{w}$ to all 3 data samples, add the new model to the plot (together with the new data point) and find the prediction for $x_t = 5$.

   (c) Repeat (b), but this time using a model without intercept term, i.e., $h_{\mathbf{w}}(x) = w_1 x$.

   (d) Repeat (b), but now using Ridge Regression with the regularization parameter $\lambda = 1$ instead of the ordinary least squares.

2. **Deriving least squares linear regression from maximum likelihood**

   Assume a linear regression model

   $$y = w_0 + w_1 x_1 + \ldots w_d x_d + \epsilon$$

   where the errors ϵ are independent, identically distributed (i.i.d.) and follow a normal distribution with zero mean $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$.

   Show that the maximum likelihood estimate for the weights in this case is equivalent to the least squares solution, i.e., to the weight optimization under the squared error loss.

3. Consider the following training data

x	y
1	3
2	1
3	0.5

   Suppose the data comes from a model $y = cx^\beta + noise$, for unknown constants c and β. Use least squares linear regression to find an estimate of c and β.

4. Consider the following data set

Item	x_1	x_2	Class y
A	1	2	yes=1
B	2	1	yes=1
C	1	1	no=0
D	1	0	no=0

   (a) Is this dataset linearly separable? Explain.

   (b) We are training a perceptron (linear classifier with a hard threshold) on this data. We append $x_0 = 1$ to each data point to account for the bias term (e.g. for item A, $(1, 2)$ becomes $(1, 1, 2)$). Suppose the current weigths to be $\mathbf{w} = (0, -1, 1)$. Assume a learning rate $\alpha = 0.2$. How should the weights be updated if point A is considered?

   (c) How should the weights be updated if point B is considered?

5. You are asked to use regularized linear regression to predict the target $y \in \mathbb{R}$ from the eight-dimensional feature vector $\mathbf{x} \in \mathbb{R}^8$. Let us define the model $y = \mathbf{w}^T\mathbf{x}$ and consider the following objective functions:

$$\min_{\mathbf{w}} \sum_{i=1}^{n} \left(y^{(i)} - \mathbf{w}^T\mathbf{x}^{(i)}\right)^2 \tag{1}$$

$$\min_{\mathbf{w}} \sum_{i=1}^{n} \left(y^{(i)} - \mathbf{w}^T\mathbf{x}^{(i)}\right)^2 + \lambda \sum_{j=1}^{8} w_j^2 \tag{2}$$

$$\min_{\mathbf{w}} \sum_{i=1}^{n} \left(y^{(i)} - \mathbf{w}^T\mathbf{x}^{(i)}\right)^2 + \lambda \sum_{j=1}^{8} |w_j| \tag{3}$$

(a) Circle regularization terms in the objective functions above.

(b) For large values of λ in objective (2) the *bias* would:

- increase
- decrease
- remain unaffected?

(c) For large values of λ in objective (3) the *variance* would:

- increase
- decrease
- remain unaffected?

(d) The following table contains the weights learned for all three objective functions (not in any particular order). Beside each objective write the appropriate column label (A, B, or C):

	Column A	Column B	Column C
w_1	0.60	0.38	0.50
w_2	0.30	0.23	0.20
w_3	-0.10	-0.02	0.00
w_4	0.20	0.15	0.09
w_5	0.30	0.21	0.00
w_6	0.20	0.03	0.00
w_7	0.02	0.04	0.00
w_8	0.26	0.12	0.05

- Objective (1): _____
- Objective (2): _____
- Objective (3): _____

6. Alice is building a model to know if a student will pass the exam or not. She has the hypothesis that the hours of study will be a good indicator of passing or not. She surveys her friends, 7 who failed the exam and 7 who passed, and asks them how many hours they spent studying. Here is what she finds:

   - Number of hours of study for each student who failed: $[1, 2, 2, 3, 5, 5, 6]$
   - Number of hours of study for each student who passed: $[5, 5, 7, 9, 9, 10, 11]$

   She trains a logistic regression classifier on the data and plots the classifier against the data, see Figure 1.

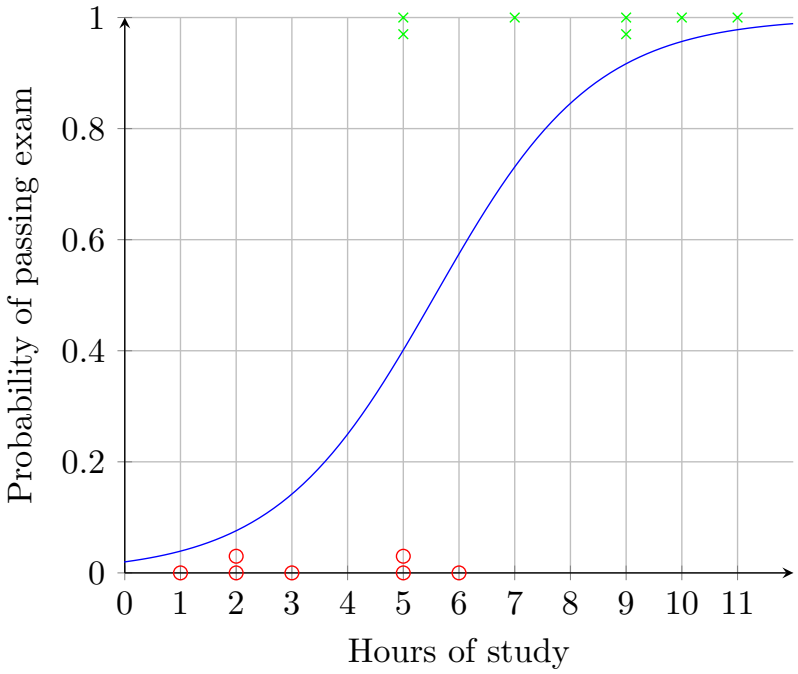


Figure 1: Plot of Alice's classifier model. The blue line represents the output of the model.

Answer the following questions assuming the logistic regression model:

(a) Consider a student who spent 5 hours studying. According to the model, what is roughly the probability that he will fail the exam? The probability that he will pass?

(b) How many hours a student must study for the model to guarantee without a doubt that she will pass?

(c) How is a logistic regression model normally turned into a binary classifier? If you turn the model into a classifier in this way, what is the accuracy of the classifier on the training data?

(d) It is most important to reduce false negatives, i.e. we want to avoid that a pass is classified as a fail. To achieve that, we want to avoid false negatives in the training dataset. How can this goal be described in terms of training precision and recall? How can the logistic regression classifier be modified to try to achieve this goal?