Lab Class ML:II

Until Wednesday, Nov. 20th, 2024, 11:59 pm CET, solutions to the following exercises must be submitted as one zip-file named `ML24-ex2-group<your-group-number>.zip` via Moodle:
1, 2, 3a,b, 4, 5,a,b and 6.

Exercise 1 : Machine Learning Basics (0.5+0.5+0.5+0.5+0.5+0.5=3 Points)

(a) Name these five concepts: x, $\mathbf{x}$, $\mathbf{X}$, X, X

(b) Give the hypothesis space H of linear regression with p features.

(c) Explain the Bayes error.

(d) How can one reduce the Bayes error?

(e) Give an example of a dataset D_1 with (label) noise: $D_1 = \{\dots\}$

(f) Given this dataset D_2, take a (class-)stratified sample $D_{2,tr}$ of D_2 with $|D_{2,tr}| = 6$:
$D_2 = \{(\mathbf{x}_1, c_1), (\mathbf{x}_2, c_2), (\mathbf{x}_3, c_3), (\mathbf{x}_4, c_2), (\mathbf{x}_5, c_2), (\mathbf{x}_6, c_3), (\mathbf{x}_7, c_1), (\mathbf{x}_8, c_3), (\mathbf{x}_9, c_2), (\mathbf{x}_{10}, c_2), (\mathbf{x}_{11}, c_3), (\mathbf{x}_{12}, c_2)\}$
$D_{2,tr} = \{\dots\}$

Exercise 2 : Probabilistic Foundation of the True Misclassification Rate (1.5+1.5+1=4 Points)

Consider a sample space $\Omega = \{o_1, o_2, o_3, o_4, o_5, o_6\}$ with six outcomes; i.e., each elementary event $\{o_i\}$ corresponds to observing one of six distinct objects. Let $\mathbf{X} \subset \mathbf{R}^2$ be a feature space, $C = \{0, 1\}$ be a set of two classes, and P be a probability measure defined on $\mathcal{P}(\Omega)$. Further, let $\mathsf{X} : \Omega \to \mathbf{X}$, and $\mathsf{C} : \Omega \to C$ be two random variables defined according to this table:

o_i	$P(\{o_i\})$	$\mathsf{X}(o_i)$	$\mathsf{C}(o_i)$
o_1	0.1	$(0,1)^T$	0
o_2	0.3	$(0,1)^T$	1
o_3	0.2	$(0,1)^T$	0
o_4	0.2	$(1,0)^T$	1
o_5	0.1	$(1,0)^T$	0
o_6	0.1	$(0,0)^T$	0

(a) Specify the joint distribution function $p(\mathbf{x}, c) := P(\mathsf{X}{=}\mathbf{x}, \mathsf{C}{=}c)$ by completing this table:

$\mathbf{x}$	c	$p(\mathbf{x}, c)$
$(0,0)^T$	0	$\dots$
$(0,0)^T$	1	$\dots$
$\vdots$		

(b) Specify the Bayes classifier $y^*()$ by completing this table (potentially more than one correct answer):

$\mathbf{x}$	$y^*(\mathbf{x})$
$(0,0)^T$	$\dots$
$(0,1)^T$	$\dots$
$\vdots$	

(c) Specify the true misclassification rate Err^* of the Bayes classifier.

Exercise 3 : Evaluating Effectiveness (2+1+0=3 Points)

Consider the following family of classification models:

$$y_\pi(\mathbf{x}) = w \cdot x_\pi$$

where $w \in \{1, -1\}$ is a model parameter learned from data, and $\pi \in \{1, \ldots, p\}$ is a hyperparameter selected manually beforehand. During training, the parameter w is chosen according to the simple learning algorithm shown on the left:

Input: Hyperparameter π and dataset D.

Output: Model Parameter w.

Learn(D, π)

1. **Initialize:** $\mathcal{L}_+ = 0$, $\mathcal{L}_- = 0$
2. **Loop:** For each example $(\mathbf{x}, c_\mathbf{x}) \in D$
   $\mathcal{L}_+ = \mathcal{L}_+ + I_{\neq}(x_\pi, c_\mathbf{x})$
   $\mathcal{L}_- = \mathcal{L}_- + I_{\neq}(-x_\pi, c_\mathbf{x})$
3. **If** $\mathcal{L}_+ \leq \mathcal{L}_-$ **Then**
   Return $w = 1$
   Else
   Return $w = -1$

Hint:
The indicator function $I_{\neq}$ is defined as in the lecturenotes slides:

$$I_{\neq}(a, b) = \begin{cases} 0 & \text{if } a = b \\ 1 & \text{otherwise} \end{cases}$$

Example:
Given
$D = \{((1, -1), 1), ((-1, 1), -1)\}$,
we get:
Learn$(D, 1) = 1$ and
Learn$(D, 2) = -1$.

You are given the following dataset D (with $p = 2$):

	x_1	x_2	c
$\mathbf{x}_1$	1	1	1
$\mathbf{x}_2$	-1	1	-1
$\mathbf{x}_3$	1	1	-1
$\mathbf{x}_4$	1	1	1
$\mathbf{x}_5$	1	1	-1
$\mathbf{x}_6$	-1	1	1
$\mathbf{x}_7$	1	1	-1
$\mathbf{x}_8$	-1	-1	-1
$\mathbf{x}_9$	1	1	1
$\mathbf{x}_{10}$	1	-1	1

(a) Let the hyperparameter π be fixed at $\pi = 1$. Using the algorithm **Learn** given above, train a classifier $y_1()$ on all of D, and determine the training error $Err(y_1(), D)$.

(b) Let $D_{\text{test}} = \{\mathbf{x}_8, \mathbf{x}_9, \mathbf{x}_{10}\}$ be the test set. Leaving $\pi = 1$ as before, train classifier $y'_1()$ on $D_{tr} = D \setminus D_{test}$ and determine the holdout error $Err(y'_1(), D_{test})$.

(c) Let $D_{val1} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\}$ and $D_{val2} = \{\mathbf{x}_5, \mathbf{x}_6, \mathbf{x}_7\}$ be the sets used for model selection with $k = 2$ validation sets (see slides). Determine π^*, and then determine the holdout error for $y_{\pi^*}()$.

Exercise 4 : Receiver Operating Characteristic (ROC) (2+2.5+2+0.5=7 Points)

Consider the binary classification scenario of spam mail detection with two classes: mail is spam ($c = 1$, "positive") and mail is not spam ($c = 0$ "negative").

(a) Look up (e.g., on Wikipedia) the following six concepts and define them for the spam mail scenario: true positive, false positive, false negative, true negative, false positive rate, true positive rate.

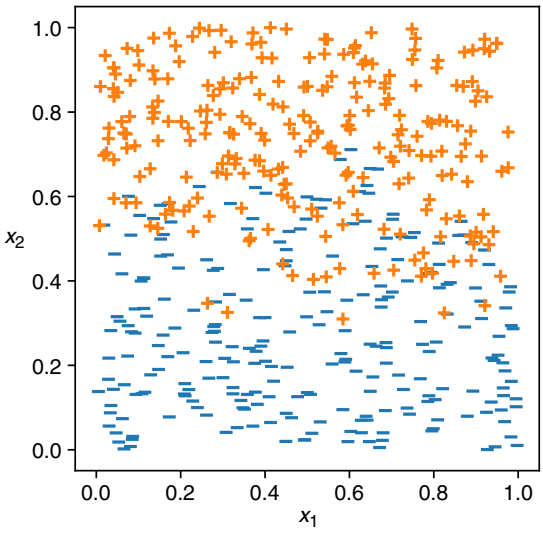
(b) Consider a large dataset D for spam mail detection with balanced class distribution, i.e., $P(C = 0) = P(C = 1) = 0.5$. Calculate the (expected) false positive rate and true positive rate for each of these classifiers:

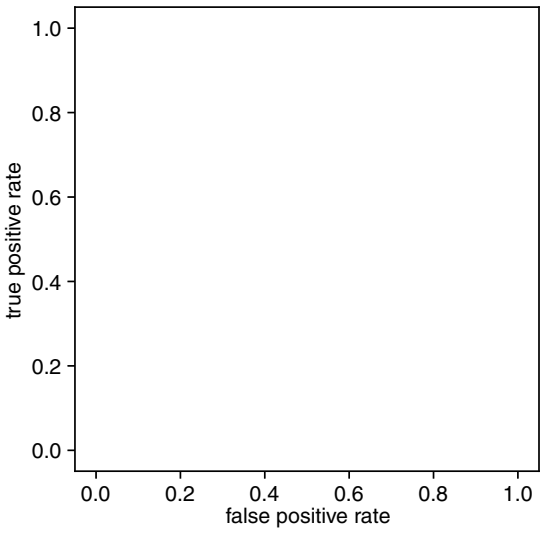
- A classifier that classifies every mail as spam
- A classifier that classifies every mail as not spam
- A classifier that classifies every mail correctly
- A classifier that classifies every mail incorrectly
- A classifier that classifies every mail randomly with equal class probability

(c) Now assume the plot on the left hand side shows D with spam examples ($c = 1$) represented as $+$ and non-spam examples ($c = 0$) represented as $-$. Consider two classifiers $y_\pi(\mathbf{x})$ for $\pi \in \{1, 2\}$ which use a threshold w_0 to classify instances solely based on either x_1 or x_2:

$$y_\pi(\mathbf{x}) = \begin{cases} 1 & \text{if } x_\pi \geq w_0 \\ 0 & \text{otherwise} \end{cases}$$

As one continuously increases w_0 from 0 to 1, both classifiers change from classifying every mail as spam to classifying no mail as spam, with the true positive rate and false positive rate changing accordingly. This continuous change of rates for a classifier corresponds to a line in the false positive rate / true positive rate scatter plot (empty plot on the right hand side), which is known as *receiver operating characteristic* (ROC) curve. Roughly sketch the ROC curves of y_1 and y_2.





(d) Based on the ROC curves you sketched in (c), which classifier do you prefer? Argue solely based on the ROC curves!

Exercise 5 : Linear Regression (2+1+0+0+0=3 Points)

This table describes four cars by their age, mileage, and stopping distance for a full braking at 100km/h:

Car	Wartburg	Moskvich	Lada	Trabi
Age (year)	5	7	15	28
Mileage (km)	30 530	90 000	159 899	270 564
Stopping distance (meter)	50	79	124	300

(a) Determine the linear regression weights w_i for predicting the stopping distance from only the age.

(b) Extrapolate the expected average stopping distance for the Lada car (i.e., age = 15 years) using the model from (a).

(c) Consider the mileage of the cars as an additional variable and repeat (a) and (b) under this setting.

(d) Draw a scatter plot of the data points, and the linear regression for a variable of your choice (i.e., either age or mileage on the x-axis).

(e) Discuss the problems and pitfalls of extrapolation.

Exercise 6 : P Basic Data Analysis and Linear Regression (1+2+1+1 Points)

For programming exercises like this one, write code in Python 3.10 or later. Submit all code that you write. You are restricted to built-in Python modules and functions, except NumPy, Pandas, and (for plotting) matplotlib or seaborn. We provide data in tab-separated-value (TSV) format – you can use the built-in csv library's `DictReader` and `DictWriter` with `delimiter='\t'` or Pandas `read_csv` and `to_csv` functions with `sep='\t'` for reading and writing.

Download and use these files from Moodle:

- `features-train.tsv`: Feature vectors for each example in the training set
- `labels-train.tsv`: Labels for each example in the training set indicating the class *is_human* ($C = \{\texttt{True}, \texttt{False}\}$)
- `features-test.tsv`: Feature vectors for each example in the test set

(a) Select two features (e.g. *num_words* and *num_characters*) and plot a scatterplot for the examples in the training set between the two features. Color the points according to the class *is_human*. Submit the plot.

(b) Implement the LMS algorithm and use it to compute the weight vector (w_0, w_1) and add the line of best fit to your plot from (a). Submit your algorithm implementation and the updated plot.

(c) Compute the residual sum of squares (RSS) for the weight vector from (b).

(d) Use the weight vector from (b) to classify each example in the test set for *is_human* ($C = \{\texttt{True}, \texttt{False}\}$). Write the predicted classes to a `predictions-test.tsv` in the same format as the `labels-train.tsv` (columns `id` and `is_human`). Submit the file with the predictions.