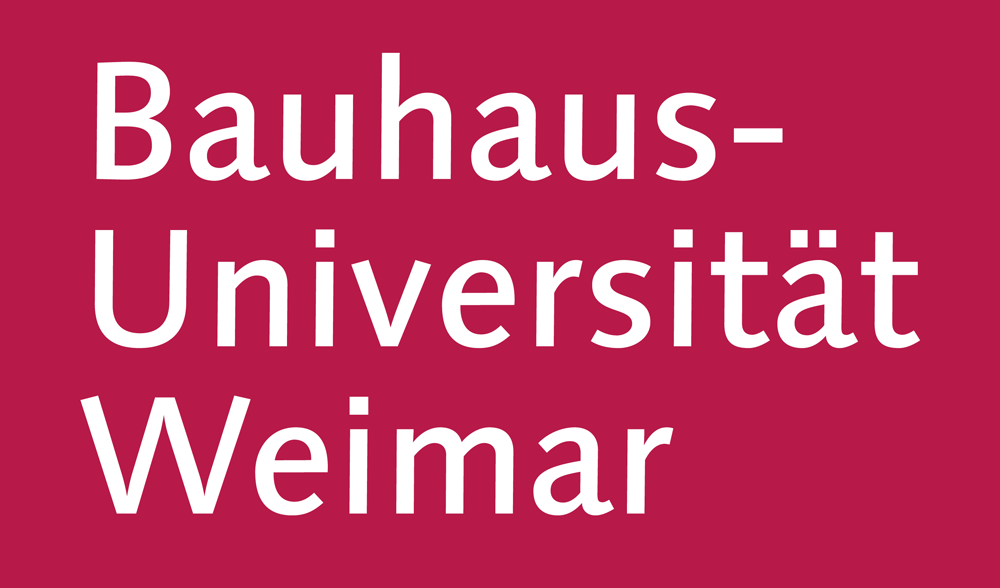
**INTRODUCTION MACHINE LEARNING**

**EXERCISE 2**



|  |
| --- |
| **TEACHER:** |
| Johannes Kiesel |

|  |
| --- |
| **GROUP:** |
| Group 16 |
|  |
| **SUBMITTED BY:** |
| Aaron Perez Herrera |
| Cesar Fernando Gamba Tiusaba |
| Chun Ting Lin |
| Olubunmi Emmanuel Ogunleye |

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

1. Name these five concepts: x, x, X, X, X

* **χ:** A single feature or variable (e.g., age, height).
* **x:** A feature vector, which represents a single data point with multiple features (e.g. (x1,x2,…,xp)) where each xi is a feature.
* **X:** A set of all feature vectors in a dataset, typically called the design matrix in a dataset, where each row represents a feature vector for one instance, in other words is the feature space, Cartesian product of the domains of the p dimensions of a feature vector x.
* **X:** Random variable (randomness regarding feature x of an object o)
* **X:** Multivariate random variable, random vector (randomness regarding feature vector x of an object.

1. **Give the hypothesis space H of linear regression with p features.**

In linear regression with ρ features, the hypothesis space Η is the set of all linear functions that can be used to model the relationship between the features and the target variable. For ρ features, this is represented as:

H={h(x)=w\_0+w\_1\*x\_1+w\_2\*x\_2+⋯〖+w〗\_p\*x\_p│w\_0,w\_1,w\_2,…,w\_p ϵ R}

This space contains all possible linear combinations of the features, parameterized by the weights w0,w1,…,wpw\_0, w\_1, \ldots, w\_pw0,w1,…,wp.

1. **Explain the Bayes error.**

The Bayes error is the lowest possible error rate for a classifier. It represents the irreducible error due to overlapping distributions in the feature space. Even with a perfect classifier, the Bayes error is the error resulting from inherent uncertainty in the data itself, where the best prediction could still be incorrect due to ambiguity in class labels for certain feature values.

1. **How can one reduce the Bayes error?**

The Bayes error cannot be eliminated, but it can be reduced by:

* Gathering more relevant features that accomplish a better separation of the classes.
* Improving data quality to ensure more distinct separability between classes.
* Increasing the sample size, which could potentially reveal patterns or distributions with less overlap if the classes are indeed separable.

1. **(e) Give an example of a dataset D1 with (label) noise: D1 = {. . .}**

The dataset is:

To take a class-stratified sample D\_(2,tr) with 6 instances, ensure the sample maintains the same proportion of classes as in the original dataset. Let’s assume each class has equal representation in this sample:

If classes c\_1, c\_2, and c\_3 have instances in D\_2, then a possible stratified sample could be:

This sample includes two instances from each class, matching the class distribution in the dataset.

**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 X subset R ^ 2 be a feature space, C = {0, 1} be of two classes, and P be a probability measure defined on {P}. Further, let X: Omega -> X and C : Omega -> C be two random variables defined according to this table:

A table with numbers and letters

Description automatically generated

1. Specify the joint distribution function p(x, c) := P (X=x, C=c) by completing this table:

|  |  |  |
| --- | --- | --- |
| X | C | Ρ(x,c) |
| (0,0)T | 0 | 0.1 |
| (0,1)T | 0 | 0.3 |
| (0,1)T | 1 | 0.3 |
| (1,0)T | 0 | 0.1 |
| (1,0)T | 1 | 0.2 |

1. Specify the Bayes classifier y\*() by completing this table (potentially more than one correct answer):

|  |  |  |
| --- | --- | --- |
| X | C | Ρ(x,c) |
| (0,0)T | 0 | 0.1 |
| (0,1)T | 0 | 0.3 |
| (1,0)T | 1 | 0.1 |

1. Specify the true misclassification rate Err ∗ of the Bayes classifier.

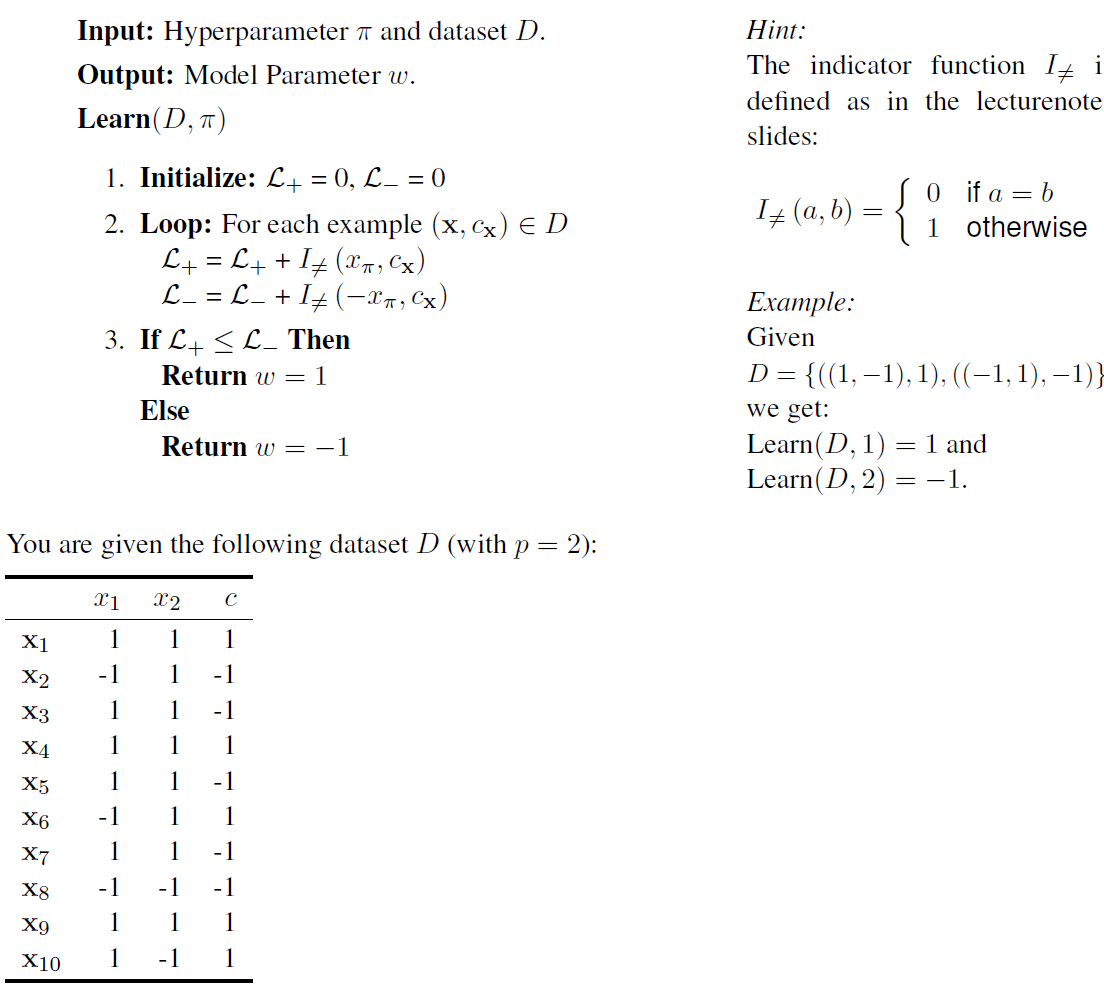
Err\* = p( X = (0,1)T, C = 1) + p( X = (1,0)T, C = 0) = 0.3 + 0.1 = 0.4

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

Consider the following family of classification models:



where w ∈ {1,−1} is a model parameter learned from data, and π ∈ {1, . . . , p} is a hyperparameter selected manually beforehand. During training, the parameter w is chosen according to the simple learning algorithm shown on the left:



With DataSet:

Function: Weight:

1. Let the hyperparameter π be fixed at π = 1. Using the algorithm Learn given above, train a classifier y1() on all of D, and determine the training error Err(y1(),D).

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X train | X1 | C |  | ζ+ | ζ- |  |  |  |
| 1 | 1 | 1 |  | 0 | 0+1=1 |  | 1 | T |
| 2 | -1 | -1 |  | 0 | 1+1=2 |  | -1 | T |
| 3 | 1 | -1 |  | 0+1=1 | 2 |  | 1 | F |
| 4 | 1 | 1 |  | 1 | 2+1=3 |  | 1 | T |
| 5 | 1 | -1 |  | 1+1=2 | 3 |  | 1 | F |
| 6 | -1 | 1 |  | 2+1=3 | 3 |  | -1 | F |
| 7 | 1 | -1 |  | 3+1=4 | 3 |  | 1 | F |
| 8 | -1 | -1 |  | 4 | 3+1=4 |  | -1 | T |
| 9 | 1 | 1 |  | 4 | 4+1=5 |  | 1 | T |
| 10 | 1 | 1 |  | 4 | 5+1=6 |  | 1 | T |
|  |  |  | TOT. | **4** | **6** |  |  |  |

ζ+ < ζ-

So the value of w = 1

1. Let Dtest = {x8, x9, x10} be the test set. Leaving π = 1 as before, train classifier y′1 () on Dtr = D \ Dtest and determine the holdout error Err(y′1 (),Dtest ).

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X train | X1 | C |  | ζ+ | ζ- |  |  |  |
| 1 | 1 | 1 |  | 0 | 0+1=1 |  |  |  |
| 2 | -1 | -1 |  | 0 | 1+1=2 |  |  |  |
| 3 | 1 | -1 |  | 0+1=1 | 2 |  |  |  |
| 4 | 1 | 1 |  | 1 | 2+1=3 |  |  |  |
| 5 | 1 | -1 |  | 1+1=2 | 3 |  |  |  |
| 6 | -1 | 1 |  | 2+1=3 | 3 |  |  |  |
| 7 | 1 | -1 |  | 3+1=4 | 3 |  |  |  |
|  |  |  | TOT. | **4** | **3** |  |  |  |

ζ+ > ζ-

So the value of w = -1

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X test | X1 | C |  |  |  |  |  |  |
| 8 | -1 | -1 |  |  |  |  | 1 | F |
| 9 | 1 | 1 |  |  |  |  | -1 | F |
| 10 | 1 | 1 |  |  |  |  | -1 | F |

1. Let Dval 1 = {x1, x2, x3, x4} and Dval 2 = {x5, x6, x7} be the sets used for model selection with k = 2 validation sets (see slides). Determine π∗, and then determine the holdout error for yπ∗().

*K = 1 ---- Dvalid = {x1,x2,x3,x4}; Dtrain = {x5,x6,x7}; Dtest = {x8,x9,x10};*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X train | X1 | C |  | ζ+ | ζ- |  |  |  |
| 5 | 1 | -1 |  | 0+1=1 | 0 |  |  |  |
| 6 | -1 | 1 |  | 1+1=2 | 0 |  |  |  |
| 7 | 1 | -1 |  | 2+1=3 | 0 |  |  |  |
|  |  |  | TOT. | 3 | 0 |  |  |  |

ζ+ > ζ-

So the value of w = -1

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X valid | X1 | C |  |  |  |  |  |  |
| 1 | 1 | 1 |  |  |  |  | -1 | F |
| 2 | -1 | -1 |  |  |  |  | 1 | F |
| 3 | 1 | -1 |  |  |  |  | -1 | T |
| 4 | 1 | 1 |  |  |  |  | -1 | F |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X test | X1 | C |  |  |  |  |  |  |
| 8 | -1 | -1 |  |  |  |  | 1 | F |
| 9 | 1 | 1 |  |  |  |  | -1 | F |
| 10 | 1 | 1 |  |  |  |  | -1 | F |

*K = 2 ---- Dvalid = {x5,x6,x7}; Dtrain = {x1,x2,x3,x4}; Dtest = {x8,x9,x10};*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X train | X1 | C |  | ζ+ | ζ- |  |  |  |
| 1 | 1 | 1 |  | 0 | 0+1=1 |  |  |  |
| 2 | -1 | -1 |  | 0 | 1+1=2 |  |  |  |
| 3 | 1 | -1 |  | 0+1=1 | 2 |  |  |  |
| 4 | 1 | 1 |  | 1 | 2+1=3 |  |  |  |
|  |  |  | TOT. | 1 | **3** |  |  |  |

ζ+ < ζ-

So the value of w = 1

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X valid | X1 | C |  |  |  |  |  |  |
| 5 | 1 | -1 |  |  |  |  | 1 | F |
| 6 | -1 | 1 |  |  |  |  | -1 | F |
| 7 | 1 | -1 |  |  |  |  | 1 | F |
|  |  |  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X test | X1 | C |  |  |  |  |  |  |
| 8 | -1 | -1 |  |  |  |  | -1 | T |
| 9 | 1 | 1 |  |  |  |  | 1 | T |
| 10 | 1 | 1 |  |  |  |  | 1 | T |

As result we have:

**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”).

1. 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.

* **True Positive (TP)**: These are cases in which a spam email is correctly identified as spam by the algorithm. In other words, when is properly send it to spam box.
* **False Positive (FP)**: These are the cases in which a normal (non-spam) email is incorrectly detected as spam by the algorithm. Here the system make user lose a possible valuable email.
* **False Negative (FN)**: This is when a spam email is incorrectly identified as non-spam by the algorithm. In this case, the system fails to detect spam and allows it into the inbox.
* **True Negative (TN)**: This is when a non-spam email is correctly identified as non-spam by the system. In other words, a legitimate email is correctly allowed to go into the inbox of the user.
* **False Positive Rate (FPR)**: This is the proportion of actual non-spam emails that are incorrectly classified as spam. It is calculated with the following formula: (Indicates how often algorithm misclassified normal emails as spam)
* **True Positive Rate (TPR)**: Also known as sensitivity or recall, this is the proportion of actual spam emails that are correctly identified as spam by the system. It is calculated as:

1. 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

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| DATA | | |  |  |  |  |  |  |  |
| TP | 1 | 100 % |  |  |  |  |  |  |  |
| FP | 1 | 100 % |  |  |  |  |  |  |  |
| FN | 0 | 0 % |  |  |  |  |  |  |  |
| TN | 0 | 0 % |  |  |  |  |  |  |  |

* A classifier that classifies every mail as not spam

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| DATA | | |  |  |  |  |  |  |  |
| TP | 0 | 0 % |  |  |  |  |  |  |  |
| FP | 0 | 0 % |  |  |  |  |  |  |  |
| FN | 1 | 100 % |  |  |  |  |  |  |  |
| TN | 1 | 100 % |  |  |  |  |  |  |  |

* A classifier that classifies every mail correctly

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| DATA | | |  |  |  |  |  |  |  |
| TP | 1 | 100 % |  |  |  |  |  |  |  |
| FP | 0 | 0 % |  |  |  |  |  |  |  |
| FN | 0 | 0 % |  |  |  |  |  |  |  |
| TN | 1 | 100 % |  |  |  |  |  |  |  |

* A classifier that classifies every mail incorrectly

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| DATA | | |  |  |  |  |  |  |  |
| TP | 0 | 100 % |  |  |  |  |  |  |  |
| FP | 1 | 100 % |  |  |  |  |  |  |  |
| FN | 1 | 100 % |  |  |  |  |  |  |  |
| TN | 0 | 100 % |  |  |  |  |  |  |  |

* A classifier that classifies every mail randomly with equal class probability

Random Classification So:

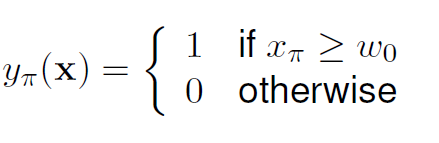
C = {0, 1} ------- Prob. C(1) = 0,5 = 50% Prob. C(0) = 0.5 = 50%

Value that is X,C(i) = Prob. C(i) \* 0,5 (0.5 = Random value 50%)

Prob (spam + true detect) = 0,5 \*0,5 = 0.25

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| DATA | | |  |  |  |  |  |  |  |
| TP | 0.5\*0.5 = 0.25 | 25 % |  |  |  |  |  |  |  |
| FP | 0.5\*0.5 = 0.25 | 25 % |  |  |  |  |  |  |  |
| FN | 0.5\*0.5 = 0.25 | 25 % |  |  |  |  |  |  |  |
| TN | 0.5\*0.5 = 0.25 | 25 % |  |  |  |  |  |  |  |

1. 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 yπ(x) for π ∈ {1, 2} which use a threshold w0 to classify instances solely based on either x1 or x2:



As one continuously increases w0 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 y1 and y2.

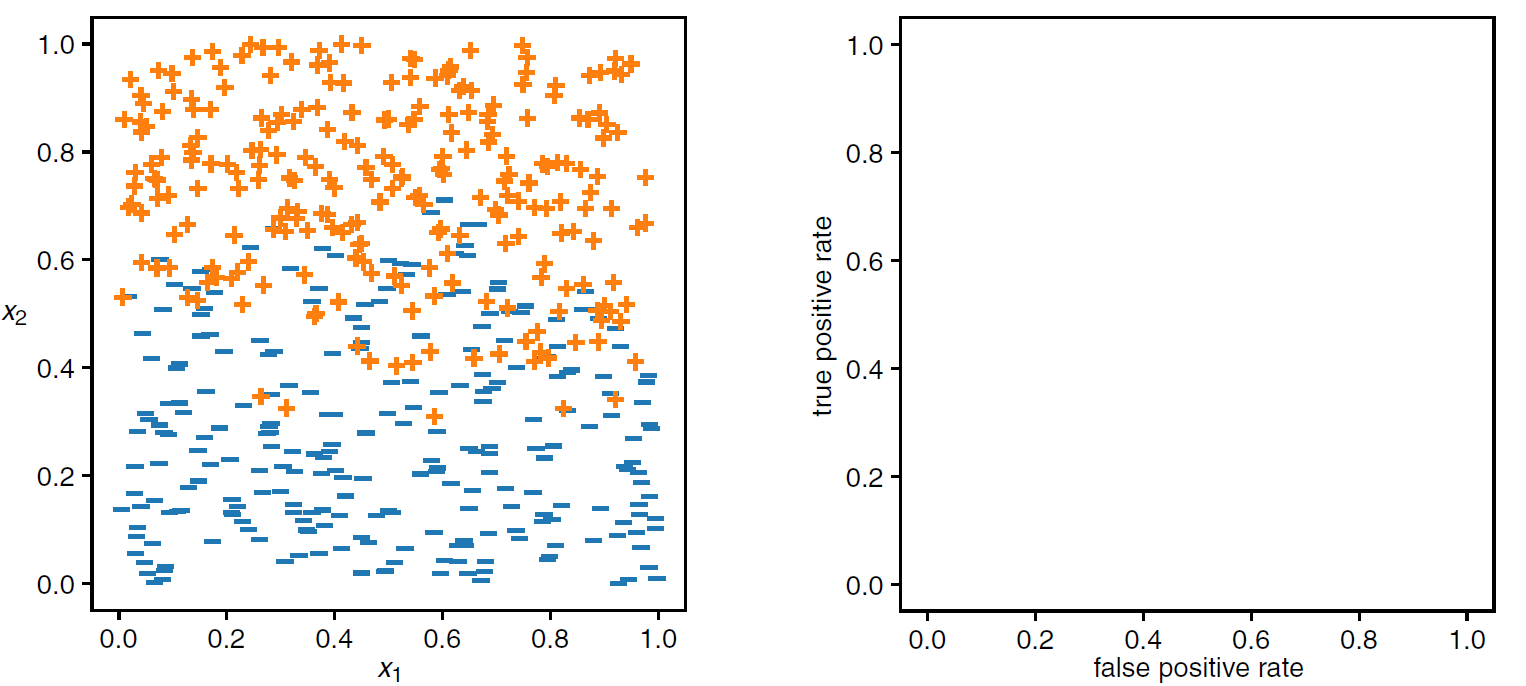
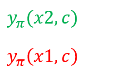
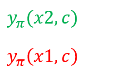
w0 = 0 ---------- all emails C=1 (Is Spam) P1 = (FPR=1,TPR=1)

w=0.25 ---------- P2(x1) = (FPR=0.8,TPR=0.8) P2(x2)= (FPR=0.5,TPR=1)

w0=0.5 ---------- P3(x1) = (FPR=0.5,TPR=0.5) P4(x2)= (FPR=0.25,TPR=0.8)

w0=0.75 ---------- P5(x1) = (FPR=0.3,TPR=0.3) P5(x2)= (FPR=0,TPR=0.6)

w0= 1 ---------- all emails C=0 (Non Spam) P6 = (FPR=0,TPR=0)



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

I would use the second classifier, because it has a curve which separates the data in a way that provides in almost each weight that can be assigned the lowest value of “FPR” and the highest value of “TPR”, thus achieving a lower error in classification rate than its counterpart (that is almost a straight-line function) and therefore a more effective classification.