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 o)

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:

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 is determined by the overlap in the class distributions, so it cannot be reduced by improving the model alone. However, it can be reduce the Bayes error by:

Gathering more informative features that better separate 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. Give an example of a dataset *D*1 with (label) noise: *D*1 = {*. . .*}

Suppose that, in this dataset, label noise has been introduced. For instance, the true label of the point might be 0, but it is mislabeled as 1.

1. Given this dataset *D*2, take a (class-)stratified sample *D*2*,tr* of *D*2 with |*D*2*,tr* | = 6:

*D*2 = {(**x**1*, c*1)*,* (**x**2*, c*2)*,* (**x**3*, c*3)*,* (**x**4*, c*2)*,* (**x**5*, c*2)*,* (**x**6*, c*3)*,* (**x**7*, c*1)*,* (**x**8*, c*3)*,* (**x**9*, c*2)*,* (**x**10*, c*2)*,* (**x**11*, c*3)*,* (**x**12*, c*2)}

*D*2*,tr* = {*. . .*}

The dataset is:

To take a class-stratified sample 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 ​, ​, and ​ have instances in ​, 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