# DD2421 Machine Learning Exams

## K-NN Classification

For 1-NN classification, the error on the TRAIN set will ALWAYS be ZERO!

For calculating the error on the test set, we need to know how much each set contributes to the average error which is the average over both sets. The formula is as follows,

So, for example, The data is split 2/3 for train and 1/3 for test, and the average error is 20%. Then the Test error for a 1-NN would be:

For a Logistic Regression model with train error of 10% and average error of 15%. The test error would be:

We would therefore choose Logistic Regression in this case as 25% < 30%

## Random Forests

There are **two principles when it comes to the randomness of the tree**. The first is the randomness of FEATURE SELECTION at each node. The second is the BOOTSTRAP REPLICATION, which is the usage of random sampling of the dataset to make the tree instead of the entire dataset.

If an estimate of a conditional probability of the forest has a value of more than 0.5, the vote of that tree can be seen as TRUE. If less, it can be seen as FALSE. If the count of TRUEs are more than FALSE, we accept that class.

For the average vote method, we will calculate the average vote of all the votes and see if the average equals more than 0.5. If the average vote is more, we will accept that class.

Diagram

Description automatically generated

## Probability Based Learning

From the above figure, we can see a probability-based learning problem, where we need to fit a distribution to a set of datapoints. Probability-based learning is all about maximizing y given datapoints x and a distribution theta.

1. For this to be a valid distribution, we need to adjust the distribution so that it fits the points.

Chart

Description automatically generated

Here we have another probability-based learning question. We need to find out which of these can be a linear regression model with normal distributed errors.

Solution: We can instantaneously disregard C as it is not a linear model. It leaves us with a) and b). A zero-mean residual means that the errors need to be equally distributed above and under the line. Meaning that the line will need to be in the middle of the points. This means that a) will not be the result of this ML estimation as the errors are not zero-mean. This means that b) is the only regression of these that may be the result of this.

A picture containing chart

Description automatically generated

## Bias-Variance

As we see above, we have some figures of testing and training errors. Generally, for models, we will see the training error slowly decrease to zero, whilst the test error will decrease and then increase as the model becomes overfitted. This means that the models will start simple, with high errors in both samples, and then increase in complexity. A simple model will have HIGH bias and LOW variance, and as the model becomes more complex, the model will have LOW bias and HIGH variance.

1. The model that is correct is 2. The reasoning is the text above.
2. When bagging, our main goal is to reduce the variance.
3. Overfitting. Read the explanation above.

Text

Description automatically generated

## Support Vector Machines

The support vector machine indicator formula is as follows:

Where alpha is a given number, t is either a 1 or -1 depending on if the support vector is positive or negative and K is the kernel function. X is a new datapoint, and S is a support vector.

In the case above we have a quadric function. We need to calculate the function for all the given points.

Text

Description automatically generated with medium confidence

## Diagram, engineering drawing Description automatically generated

## Support Vector Machines and Alpha Values

The alpha value is the relative importance of the datapoint to the classification boundaries. Meaning, a higher alpha will correspond to a more important datapoint, and vice-versa. We also know that the sum of all t \* a = 0.

1. From the answers below, we know that d = 0 and b = 4. We also know that the sum of all the alphas and their sign must be 0. This forces a to be equal to 1, otherwise the alphas would not sum up to 0.
2. The positive sample (1,1) is very close to the boundary. This means that the alpha value is VERY HIGH. 4.
3. The negative sample (0, -1) is close, but not as close as (2,1). However, it is closer than (0, -2) and therefore it gets a MEDIUM LOW value of 1.5.
4. The negative sample (0, -2) is far away from the boundary and has relatively little importance to the boundary. This means that the alpha value is LOW. 0
5. The negative sample (2,1) is close to the boundary, it will therefore have a higher value than C, so it gets 3.5.

## Maximum a Posteriori Classifiers

A MAP classifier tries to maximize the following probability:

, we can assume Pr(D) to be constant because each datapoint is equally distributed. This means that we need to maximize the numerator.

We’re trying to find the best distribution that fits our Dataset to predict new points x.

Text, letter

Description automatically generated

## Probability Based Learning and Maximum Likelihood

It is very similar to MAP, the only difference here is that we can make the fundamental assumption of I.I.D which means that every datapoint X is independent and identically distributed. This allows us to directly multiply their probabilities with each other.

Doing so, we can define ML as:

Thus, the ML estimation for a given distribution would be (example Gaussian):

This means that we need to find the that maximizes the probability.

A useful derivation of mean and variance to know is the following:

Text, letter

Description automatically generated

A picture containing text, clock, gauge

Description automatically generated

It says that the mean is just the average of all datapoints, and the variance to be the average of the square errors (Mean Squared Error, MSE)!

Table

Description automatically generated with low confidence

1. Using ML, find the probability of head. Meaning we want to:

Trial 1. We have 1 head out of 10 outcomes. = 1/10

Trial 2. We have 4 heads out of 10 outcomes. = 4/10

Trial 3. We have 3 heads out of 10 outcomes = 3/10

Trial 4. We have 2 heads out of 10 outcomes. = 2/10

It is a binary outcome which is a Bernoulli Distribution.

Text, letter

Description automatically generated

As we see from the derivation above, the ML estimation of the probability is simply the occurrences divided by the number of outcomes.

1. The new ML estimate will be the old one adjusted for new data.

Lamda = n + m / N + M

m is unknown, but can be calculated as lambda = 0.2 = m/M = m/90 = 18

Therefore, the new trials will be:

Trial 1. 18+1/100

Trial 2. 18+4/100

Trial 3. 18+3/100

Trial 4. 20/100

1. We can measure reliability as the amount of variance. A lower variance has a higher reliability. From this, we can conclude that B would be more reliable.

## Dimensionality Reduction

Dimensionality can be reduced using the PCA to reduce variables. We want to maximize the projection length of each variable. The projection length is calculated with the following formula:

A screenshot of a computer

Description automatically generated with medium confidence

1. It is the formula above.
2. We want to choose the class which maximizes the projection length of vector x.

## Information Contents

We measure information using information gain and entropy. Entropy is the measure of uncertainty in a system. Higher = More uncertainty of an outcome, and lower is vice versa. Information gain is how much better we can make the next guess, based on a new piece of evidence or information.

Shannon’s Entropy is defined as:

Text, letter

Description automatically generated

1. To answer this question, we need to know the possible outcomes of each class. We draw and place back. Each card draw can be seen independent from each other. We have a 13/52 chance of drawing our suit

The information content in bits is therefore the Shannon Entropy which can be measured as:

This is for one class. We need to calculate these for all 4 suits. It will therefore be 4 \* 0.5 which is 2 bits.

1. Firstly, what are the outcomes? Well we have 2^3 outcomes. 2 of these win. It will therefore be -2/8 \* log2(2/8) + -6/8 \* log2(6/8) = 0.811.
2. The above is the predictability of a win, meaning three cards. We know that one of the cards gives 0 entropy. We can “subtract” that cards entropy. It would be the same as subtracting 0.5 which results in 0.311….. This is not the proper explanation but it works…

## Probabilistic Linear Regression

We want to optimize the best linear regression to the points. The slope of the line is determined by the weights vector. The error of the slope is to be assumed Gaussian with zero mean and variance sigma squared.

The formula is therefore:

The posterior for y given x is a Gaussian distribution:

The free parameters here to optimize is therefore the weights as we assume the variance sigma squared to be given.

Chart, scatter chart

Description automatically generated

1. The formula can be seen above. The free variables are the weights.
2. We can calculate them by doing an ML estimation.

Why is this? Optimizing a linear regression is the same as minimizing its errors. Therefore, we want to minimize epsilon. Since epsilon is normally distributed with 0 mean, we can rewrite is as following:

Diagram, text

Description automatically generated

We notice that the only thing we really need to minimize is the (y – w \* x) which is the square error. So, the entire expression just becomes minimizing the sum of the square errors.

We will notice that the w is unknown, but to minimize we need to find the w where derivative of SSE is 0. This will be the partial derivative with respect to w0 and w1.

The derivative for this problem will be 5w0 – 10 for w0 and 10w1 – 6 for w1.

w0 = 2, w1 = 3/5

## BackProp Learning