Written by:

Ceesay, Cindy Madeleine Svendsen

Raya, Athena

Vaidya, Rujuta

Workshop 1

Introduction:

As we learned in class today, under supervised learning, we have a goal i.e. the possible output is already known. That is, we want to teach the machine the conclusions it should come up with. According to Langerød, in supervised learning we take a set of labelled data and use the knowledge acquired from this data to label new data that is unknown. For this to work according to Langerød, the desired output needs to be known and the data algorithms use to train has to has to be labelled (Langerød, 2016). Supervised learning is divided in regression and classification.

In regression : You are for example trying to predict the y variable (independent variable) based on some x variables. This could be for example trying to predict the sales of a particular dress based on x variables like the weather, fashion trends or price of the dress. There could be many more variables that affect this. We know for example that there is a negative correlation between price and sales for normal goods. This can be done using linear regression if one has quantifiable data on all the variables.

As mentioned in class, while regression is for numbers, classification is for non-quantifiable data. In classification we order and categorize data based on different categories. While the input is not determined or counted, the output must have an label (Langerød, 2016). An example is to classify a mail as either important or spam, sent or received. A program can

also be taught to recognize an animal after it has been shown a set of animals that are correctly labeled with their species.

Under unsupervised, there is no “grounded truth” or particular goals, we are extracting information from the data in order to find it’s structures. Unsupervised learning is divided in clustering, neural networks and anomaly detection. Clustering is to group unlabeled data to identify patterns. There does not need to be a training set in the unsupervised data. We cluster or group together entities such as products, animals or images together. This way we can see the differences between the different groups.

Question 1: Which of the previous challenges are supervised or unsupervised learning?

Where to place what in the grocery shop : Supervised learning, because it is possible to model that people that bought this type of product will also pick that type of product. This can be performed by an simple regression (OLS) i.e. we can find the correlations between if a person bought product X and product Y then they are more likely to for example buy product. In OLS we can also classify people as under 21, between 22 and 35 and so forth and then run a regression on sales of products when they are placed in different places in the store.

How to predict the weather: Supervised learning, as the weather are predicted by collecting quantitative data. The data can be plotted in an coordinate-system, which you later on can draw a line between the plots to estimate the weather.

Probability of an transaction being computer fraud: unsupervised learning because we are trying to find hidden patterns in the way in which fraud occurs. The patterns are unknown so we need to use unsupervised learning.

100 year flood: Unsupervised learning, we are trying to predict extreme rainfall patterns. The model does not already exist and cannot be made easily. The result depends on several inputs i.e. amount of real-time rainfall, river stage and consequences of storms and the river characteristics basin.

Testing hypothesis of clinical trials, for instance drugs or medicine: Supervised, because you can make a model or design an experiment and can predict effects based on such a model. You can give one group the placebo pill and the other the actual pill and test the effect of the pill on the people. The model can then be used for prediction.

Stock market: Unsupervised learning. This is because there are too many inputs. Usually, it is assumed that the information is already incorporated and that it is impossible to make abnormal profits in the stock market but this is not true in real life. There are outliers and people (arbitrageurs for example) making abnormal profits. This is however not predictable and cannot be supervised learning. There is unpredictability in arbitrage. This cannot be taught to a machine using supervised learning.

How many cashiers to use and predicting demand: Supervised learning. We believe that it is possible to predict this using models. You can try to figure out using patterns of number of customers and amount of time each cashier uses to check out a customer to predict how many cashiers are needed. You can also use data from other stores in the area to try and predict how many people will come to the store, and thereby predict how much staff is required at any given time. Rush hour data, traffic in the area, whether the area is residential or industrial will also help us predict this sort of thing.

Aging: supervised for natural and unsupervised unnatural death. Since the question asks us how likely a certain person is to age to age 70, we found that there are two ways of looking at this question. Either, people live to their natural lifespans or they die because of unnatural reasons like accidents. For predicting the natural lifespan we can argue for supervised learning because we can given the person's place of birth (an Indian child born in India vs an Indian child born in Norway), where they have grown up, genes, nutrition at birth and other factors technically can predict how long a person can live (naturally). We cannot however predict unnatural causes of death such as traffic accidents or wars. It can be predicted to a certain extent but not for all ways in which a person could die. So there are some ground truths

Gun purchase/ How much time it goes from a person buy a weapon until he shoots: We believe that it is unsupervised because we do not have any ground truths or particular prediction goals. It is also difficult to model using regression or classify.
Grouping persons: Unsupervised, as the computer cluster people and finds patterns.

The solar flares: Unsupervised, as there is no data from that far back in time i.e. the last solar flare and therefore this is unsupervised. The machine clusters together patterns for example to figure out similar trends as the ones people might have written down about or other tracers that we have about this phenomenon.

Question 2: Which of the supervised cases are regression or classifiers?

How to predict the weather: We believe that this is regression. A model can be built to forecast the weather given certain inputs which we get from our satellites.

Testing hypothesis of clinical trials, for instance drugs or medicine: This is also regression. We are trying to do causal analysis to find the effect of the drug.

How many cashiers to use and predicting demand: Regression since we believe that this can be modelled.

Aging:We believe that this one is regression too. We can make a model about it and it is quantifiable.

Question 3: What types of unsupervised problems are in there?

As mentioned in the introduction, the unsupervised learning is divided into clustering, neural networks and anomaly detection.

Probability of an transaction being computer fraud: Clustering can help here because hopefully the machine will be able to learn to cluster all the normal transactions and then pick out the anomaly. Or anomaly detection, whereby it clusters all the other transactions.

100 year flood: Either one of the other two methods might need to be used here. Neural networks might work because there are so many inputs that need to be covered at the same time.

Stock market: Clustering might work here, since we can cluster the stocks that react normally to market forces and new information. Arbitrageurs can then try to find the anomalies that stand out.

Aging for unnatural death: Neural network might be the best method here. There are many inputs that need to be analyzed at the same time to predict when a person will die.

Gun purchase: Clustering and anomaly detection. We could try to cluster on aspects with the purchasers and this way try to find the ones that might have a high likelihood of murdering someone.

Grouping persons: Clustering. Can see how these certain groups act.

The solar flares: Clustering, as we are trying to find a pattern based on similar things happening like before.

Question 4: Provide other examples not mentioned

1. Earthquakes in certain regions.
2. Estimation of which persons that are going to get approved their scholarship
3. Estimate the effect of natural resources on current prices / supply and demand.
4. Estimate the quantity of a product need for stores
5. Predict which phone service plan is the best for you
6. Estimate the cost of living wages per region against living expenses

Workshop 2

Question 5: Based on the previous exercise, brainstorm other kinds of predictions/classifications, or any other analytical techniques that you think may help to solve a particular problem you are aware of (make a list!)

Earthquakes in certain regions: Unsupervised, because several inputs gives one certain answer. You use a neural network in order to build a model of the quantities of the earthquakes. Beforehand you can not have an assumption of what the answer probably should be, so you need to check if you have used the right methods of research.

Estimation of which persons that are going to get approved their scholarship: Supervised, because we can access the data beforehand. Based on a binary linear research we would get a yes or no answer. You have a ground truth and have particular prediction goals that you either get approved for scholarship or not.

Estimate the effect of natural resources on current prices / supply and demand: Unsupervised, as there are too many variables involved to predict an exact answer. The result depend on i.e political issues and climate change. Neural networks are involved as several inputs give one output.

Estimate the quantity of a product need for stores: Supervised. You predict how a human behave by social science and thereby one input will give one output. Regression testing, because the machine can learn from it's past experience.

Predict which phone service plans is the best for you: Supervised learning as persons are using their phone approximately the same amount each month and approximately to the same activities. Regression testing, because the machine can learn from the users past experience.

Estimate the cost of living wages per region against living expenses: Supervision learning as . the cost will go up and down based on inflation and interest rates. The house market moves according to demand patterns, demography and wages. So as we can see there is a feedback loop between wages and house demand. There are obviously many omitted variables here, so the house market is unpredictable. Based on linear regression we can plot the variables into a coordinate system.

Question 6: What is the impact or relevance if you manage to solve the problem?

We have chosen to concentrate on one of the problems above, i.e. estimating the cost of living wages per region against living expenses i.e. we can predict whether or not the amount of minimum income earn is sufficient for cost of living per person. The relevance is that data will determined whether firms a region/city are paying high enough standard wages in comparison to cost of living.

The impact is that there will be fewer people who will earn lower than the basic income needed to live a sufficient standard of life. The government can predict faster their result of their actions. They can make an action plan and thereby increase the standard of living. It will be easier to for example see the effect of increasing interest rates on people in a certain region with this sort of data available. If one can calculate the minimum cost of living, the standard wage the person should have then the effects of the increase or decrease in interest rates per

region can be more easily identified. This way the government can try and find ways to help the regions that need it the most.

The companies can predict the demand based on their customers income, so they can give a better price for their user group. Companies can then more easily price differentiate between customers. For example they can offer some products to people with lower wages and some other products to people with higher wages. This need not be a bad thing. It means that all customers can get the product and the total welfare in the society will increase.

Question 7: What are potential "side-effects" or "dark side" of your solution, from ethical design or privacy perspectives?

If we find an area of the country where people are being paid less than they should be to have a reasonable standard of living then several companies can form a union/companionship where they cooperate to bring down the salary in other parts of the country, if they know what other companies pay for other people.

Data may reveal consistently low wages in a particular region, bringing awareness to the low income worker. The data could bring about an uproar of demand for higher wages and workers may leave their jobs. Small business owner may be forced to pay more in wages even if they can not afford to pay. The small business may have to increase price to afford products and employee wages or they might have to layoff these workers making the situation even worse.

References:

Langerød, K. (2016). *Category Clustering: Exploring feature creation and similarity.* Oslo: UiO: Department of Informatics Faculty of mathematics and natural sciences.