Week 1 datascience (sv)

4 V’s: Volume, Velocity, Variety, Veracity (DATA)

Week 2 datascience (sv)

Task (T)

Performance (P)

Experience (E)

Tasks:



Supervised learning is where you have a ground truth.

Clustering is where some lookalikes are put together.

Association: which products are bought together.

Semi-supervised learning: limited ground truth for a small part of data (Meestal)

Reinforced learning: awarding the model and minimize penalties, no ground truth.

**Regression models:** (estimate the relationship between a target variable and features.)

1.1 Linear regression: Describe an **indepent/response variable** using **explanatory variables**. Y=ax+b

1. DATA
2. MODEL
3. TRAIN
4. EVALUATE

1.2 Multivariate Regression: better fit than lin regr. Same model, different data.

Less loss is better!

1.3 Polynomial Regression: when a lineair model isn’t good fit: add x^2, x^3

Bij polynomial regression is het erg belangrijk om je data te schalen, omdat je dan 100^6 heb straks en je dus je data moet schalen met een log bijv. anders heb je te grote getallen.

**Classification models:** 0 and 1**:** (predict class) of diabetes type 1,2 of geen (dus 3 keuzes)

1.4 Logistic Regression: decision boundary, maximize P(y=1\X1) and P(y=0\X0). 0 of 1 maar wel met lijn ertussen. Tussen 1 en 0. Wanneer de kans(p) groter is dan 0,5 is het 1, anders 0.

1.5 K-Nearest Neighbours: decision boundary, it takes the class of the closest neighbour. Continuous.

1.6 Support Vector Machine: Decision boundary, maximum towards support vectors.(same as logit)

1.7 Decision Tree: Decision boundary, most purely split the classes, combines criteria to find the best tree. Elke lijn is een decision!

Week 3 datascience (sv) (model estimation)

Content:

* Machine learning hypothesis
* Loss function
* Vectorization
* Analytical approach
* Rule and gradient descent
* SK learn

Hypothese:

**Y^** = ax+b of ookwel: h(ϴ) = ϴ1 \* x + ϴ0

Minimize the loss function (**cost function**) **(J(ϴ)) = SOM afstand van de optimale lijn in het kwadraat (residuel error^2).**

J(ϴ) is altijd parabool doordat het een kwadraat functie is.

Minimize by: dJ/ dϴ = 0 afgeleide=0

**Vectorization**: vector maken van de x en ϴ -> y^ = X \* ϴ (much more efficient)

Solving the cost function: or **analytical** (fast when features <100) or **gradient descent**.

**Gradient descent**: iteratively updating ϴ: gradient is de graad van de lijn(stijlheid). Je moet een goede **learning rate ( a** ) vinden. Dit is de grootte van de stapjes die je maakt naar het minimum toe. Je hebt plus en min en daar tussen minimum, als de gradient descent dus 0,1 is is ie er bijna en -0,1 ook. Uiteindelijk overschiet de gradient descent het minimum.

Een cost function kan local minimum hebben doordat y ook een factor is.

Goede learning rate vinden door the rule of thumb(x3): 1e-5, 3e-5, 1e-4, 3e-4 etc

2.3 Multivariate linear regression :

2.5 polynomial regression: (same)

2.6 logistic regression:

Regression=MSE

Classification=percentage correct: 1 is true en 0 is false, zo dicht mogelijk tegen 1 aan is goed.

Week 3 Workshop (research methods)

* **Library** (literatuur studie) = to get an overview of the guidelines and theories
* **Field** (problem analyse) = to explore the application context
* **Workshop** (prototype) = to explore opportunities
* **Lab** (data analyse) = to test a concept of your product
* **Showroom** ( pitch ) = to test in relation to existing work.

Fit vs Exspertise

Overview vs Certainty

1. Ask yourself a question
2. Do something
3. Find a solution

Week 4 Workshop (research methods)

Knowledge – experiment – experience – observe – knowledge…(loop)

The research question: domain, reason, strategy, answer

The research question must be consistent!!

Research question must be: (SMART)

* Controllable (open in what and why)
* Skillfully (effective, efficient and permissible)
* Logical (reason correct)
* Valid (measuring the right thing)
* Reliable (it cannot be a coincidence)
* Adequate (answer the right question)

Week 4 datascience (sv) (training and validation)

* Underfitting and overfitting: what is bias and variance
* Diagnostics
* Remedies for High Variance
* Remedies for High Bias
* Hyperparameters
* Quiz: Fix problems

We want low bias and low variance. ***Low bias is in de roos*** en high bias is naast de roos. Low variance is op elkaar en high variance is uit elkaar.

X is een parameter: the number of parameters can be found by looking at the validation data!

Errors: **Bias and Variance**

Bias: the class of models is unable to fit the data. So how much the true values differs from the ‘best possible prediction’. High bias is te makkelijk model.

Variance: the class of models could fit the data but it doesn’t because parameters are hard to optimize. The variance of the mean over different systems. High variance is te moeilijk model.

**Overfitting**: High variance!

**Underfitting:** High bias!

Causes underfitting:

* Model too simple
* Not enough features
* Not trained the model enough

Causes ***overfitting***:

* ***Model too complex***
* ***Too many features: rule of thumb zegt dat je het aantal pictures moet delen door 10.***
* ***Not enough training data***
* ***Learning on a poor sample***
* ***Model is overtrained.***

What is the case when the model overfits? – the model doesn’t generalize.  
diseases: non-converge: geen minimum vinden, underfitting, overfitting, analysis of errors: rmse?

Wanneer de validation error van model 1 groter is dan van model 2 betekent dat dat model 1 overfit.

Je wil het minimum van de test of validation set!

**cross validation:** strategy to diagnose machine learning; train and test/validation set.

**Learning curve: y=loss, x=model complexity(1st order to 4th order polynomial)**

**Gradient descent:** we iteratively update theta using an update rule.. alpha controls the size of the steps.

**Oscillation:** op en neer op een minima door een te hoge alpha. ***Dus lower the learning rate! Of scale data!***

Remedy to high variance:

* More training samples
* Use less features
* Use more regularization: zorgt ervoor dat de grootte van de parameters kleiner worden.
* Early termination

***Remedy to high bias:***

* ***Use more features***
* ***Train for more iterations***
* ***Use less regularization***
* ***Increase learning rate***

Intability during training:

* Oscillation: scale your data
* Degrade after minimum: see high variance
* Non convergence: lower the learning rate

**Hyperparameters:** is a  parameter whose value is used to control the learning process.

When we plot the loss over epochs we want a learning curve that goes down really fast and stabilizes. Als er rinkeltjes in zitten overschiet hij het minimum steeds. Dan kunnen we de learning rate lowen. Or normalise data.

You should always diagnose again after you change a setting.

**Early termination** wanneer je het minimum hebt bereikt, maar daarna weer overfit. Or reduce features.

Week 5 datascience (sv) (evaluation methodology)

* Evaluation
* Cross Validation
* Evaluation Metrics
* Experiment setup
* Baseline, Hypothesis testing
* Statistical significance
* Conclusions

**Validation:** did we learn the model correctly? 🡪 use loss

**Evaluation:** did we learn the correct model? 🡪 is a metric that our audience understands

**Regression:** estimate the relationship between a target variable and features. MSE, RMSE, MAE(logistic problem) (MEAN ABSOLUTE ERROR), R2 (Het kwadraat van de correlatiecoëfficiënt),

* A model is learned to find the shortest route, the model is to be used to minimize the total mileage of all routes driven on a single day. What metric would you use?

**Classification:** predict the most likely class. **Confusion matrix** (predited guilty and is innocent= false positive), **accuracy** (the fraction of cases that was classified correctly(TP+TN)/100), **recall** (the fraction of positive cases that was correctly identified(TP/(TP+FN))), **precision** (the fraction of identified cases that was correctly identified(TP/(TP+FP)))

* What is a false positive when we predict if a student will pass their exam?
* Lets take the case of diagnosing which people have covid right now. To evaluate a model that does that, accuracy may not be very useful because..
* We wish to optimize the covid diagnosis model for ***a low number of false negatives***, what is the metric to use here? ***Precision***

In een recall/precision grafiek is degene met meeste oppervlak het best.

So:

Rules of thumb for sound machine learning:

* Check the quality of your **data**
* Law of parsimony (**model**): the simplest model is the best!
* Learn just right: **validation**
* **Evaluation:** measure **effectiveness** and compare against **baseline**
* What is an example of occams razor in machine learning? A simple model is better than overcomplicated

**Cross validation:** How good does a model generalize to unseen data?

* Random split data in training and validation set
* Is biased, but not with test set erbij

**n-fold cross validation:**

* Split the collection in n-folds
* Hold n experiments: use 1 fold as test set, remainder as training data
* Average the evaluation metrics over these n-experiments
* When should you use cross validation? Always!

**Training**: to train the model

**Validation**: tune hyper parameters and choose best configuration(welke bouwstenen zijn het best voor dit model?)

**Test**: for an unbiased estimation on future data

* The error on a validation set may underestimate the true error, it is a **biased estimation**! Why is that? C

Hyperparameter is voor je experiment vastgelegd. Je hebt meestal meerdere.

To test two models which is better? Statistical significance:

Two types of hypothesis:

H0: null hypythese is usually that the observed phenomenon is a result of chance (hoogstwaarschijnlijk) vb.: model A is as effective as model B

H1: alternatieve hypothese: result of a non-random cause (niet hoogstwaarschijnlijk) vb.: model A is not as effective as model B

Tests to figure this out:

t-test, Wilcoxon signed rank, chi squared.

***p-value<0.05***

* which system is better? A yes!

Week 6 datascience (sv) (data)

Training loss and validation loss

Content:

* cleaning
* cathegorial data
* splitting data
* scaling
* balancing
* outliers
* synthetic data

tools for processing:

* numpy
* pandas
* python

cleaning: handle missing values = remove and impute

Never impute with random value!

most models need numerical data

binary data: 1 of 0

non binary data: maandag = 1, dinsdag = 2 (dummievariables)

deze non binary data kan je met een n-1 aantal dummie kolommen omschrijven! Want als het geen maandag is, dan is het dinsdag. Het ligt aan collinearity, als meerdere features dependant zijn.

n-1 boolean variabelen heb je dus nodig!

split data in x and y

when the features are on different scales: **scaling**

* learning may be slow, oscillation
* model may not fully converge
* numerical instability

standardscalr zet de gemiddelde x naar 0 en de standaardvariatie naar 1: z-distribution

when using scaling u should want the transformation to be the same, so same mean and variance.so scale everything the same way.

**balancing:** every class has an equal proportion: when the classes differ greatly in this! So with covid u have 1% that is positive, but you want the model to find those, so you multiply that case 100 times.

**recall scores**: geeft aan hoeveel procent positief test, 80% is correct geidentificeerd als positief(heeft een leverziekte)

**precision scores**: geeft aan hoeveel van die positief geteste mensen nou ook echt positief zijn. 40% heeft geen leverziekte. Dit w

an **outlier** is an error. Using normal distribution you can find them.

* You should be careful when u extrapolate the model (extrapolate is aanpassen)

Remove values beyond 3 SD(standaarddeviatie(97%)) from mean.

Synthetic data (If datasets are small): create new data points, they don’t exist, but we are going to imagine they do. Create datapoint between two other datapoints.

Week 7 datascience (sv) (features)

Contents:

* features
* low variance
* filter method
* collinearity
* wrapper method
* embedded method
* random forest
* text representations
* time series

often only a small part of features is critical.

High dimension means a lot of features.

Possible redundancy (overtolligheid) with target variable, this can result in overfitting.

U need to find a minimal subset of features that provide optimal results:

* ***remove low variance***: als een variabele scheef is is er weinig informatie(veel vrouwen en weinig mannen), dan overfit het model. Solution: remove boolean variables with a low variance. Dus hier mannen en vrouwen 50/50 maken. With variance threshold. You should tune the threshold, so 0,8 is veel van je data dat weg gaat.
* **filter method** : ranks the features according to distance, information, dependency, consistency etc. pearsons correlation bijvoorbeeld.
* ***wrapper method***: forward selection, backward elimination. To find optimal features. Forward selection is that u start with best feature, then next best feature etc. backward is that u start with all features and then find the worst feature etc.
* with backward elimination we choose the feature that changes the most on the validation set!

**collinearity:** estimate a persons weight based on their height, feet and meters hold the same information. Solution: from two highly dependent features keep only 1 using **correlation matrix**, **VIF** removes features with high multicollinearity, look at **T-test** to see if features make significant difference. Als de correlatie bijna 1 is, is ie collinear. Trouble converging. A small change can result in a different model. VIF hoger is dan 5, is er coll. F-value geeft aan hoe significant een feature is, groter dan 0,05 is coll.

Multi collinearity : a feature is collinear with a set of other features. Inkomen en salery, rent, interest.

***Regularization***helps prevent overfitting(door penalties(lambda) toe te voegen an het model!). Any measure that we can do to prevent overfitting. The amount of regularization can be determent by lambda! Als lambda groot is passen we meer regularization toe. Als het model overfit heeft het grote getallen dus moeten we penalties toepassen om het model te straffen door lambda groot te maken!

**Random forest:** create large subset of decision trees (trained on random subset of training set, with random features) and take average of all. (ensemble method)

Ensemble methods: meerdere modellen in een 1 model

**Time series**: in time series de data is captured at equal intervals and each successive data point in the series depends on its past values.

Feature engineering: nieuwe ml toepassen. text classification for example.

Week 7 workshop (visualiseren)

Exploratory and explanatory

* nominale schaal: male/female, no order
* ordinale schaal: bad/good/best, in order
* interval schaal: time/temp, order, gemiddelde
* ratio schaal: lengte, gemiddelde, natuurlijke 0 waarde

***Oefentoets machine learning***

1. c
2. a
3. a
4. d
5. a
6. b
7. b
8. a
9. c?
10. a
11. b
12. c
13. a
14. c
15. c

post nl hoef je niet perse te noemen in het paper.

Data ook visualiseren? Kan.

Bijlagen vallen erbuiten, dan refereren naar appendix a1

Grafiek resultaten kleiner gedeelte laten zien,

Let op significantie bij rsme.

Maak een kolom voor geschaald en niet geschaald. Wrm is het beter bij geschaalde data?

Dat hebben we zo geleerd

Geen signigifanct verschil, niet geschaald in bijlage.

Een lijn bij grafiek van resultaten die aangeeft dat het pasen is.

Sarima: zet de punten erbij, niet alleen lijn.

Data en predictions veranderen!!

Benoem de geschaalde data! Zorg dat het duidelijk is wat op de y as staat

Residu plot erbij!

Voorspelling set is test set.

Rolling window validation zie je verschuiving, hier ga je vragen over krijgen!!