Mechatronic Engineering program:
Python for machine learning and data science

Data and models interpretation

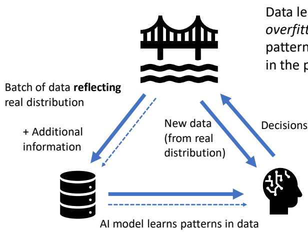
Ziemowit Dworakowski
AGH University of Krakow

1

Data leakage

Data leakage occurs, when the decision system (*e.g. a classifier*) uses information during training and initial testing that will not be present in-operation.

Data leakage can be viewed also as a „*hidden overfitting*” phenomenon: A model learns patterns that are not general but present only in the particular setup of training data.



2

Data leakage

Data leakage can happen due to:

- Training example leakage (row-wise leakage) – training samples influencing directly testing samples (poor augmentation or data acquisition)
- Feature leakage (column-wise leakage) – inclusion of columns directly related to targets
- Data acquisition procedures that differ for training data and in-operation data
- Lack of independence of samples – e.g. by indirect inclusion of time-related information

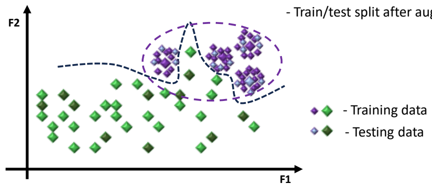
3

Data leakage

Data leakage can happen due to:

Training example leakage (row-wise leakage) – training samples influencing directly testing samples (poor augmentation or data acquisition)

- Train/test split after augmentation of data



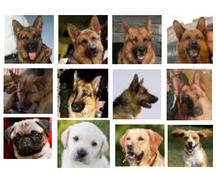
4

Data leakage

Data leakage can happen due to:

Training example leakage (row-wise leakage) – training samples influencing directly testing samples (poor augmentation or data acquisition)

- Train/test split after augmentation of data
- Samples being „copies" of each other



These are all pictures of the same dog!

Even if we add different dogs to the dataset, pictures of the same one should not be split into training and testing at the same time!

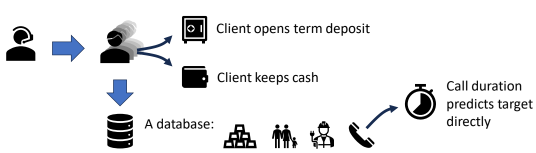
5

Data leakage

Data leakage can happen due to:

Training example leakage (row-wise leakage) – training samples influencing directly testing samples (poor augmentation or data acquisition)

Feature leakage (column-wise leakage) – inclusion of columns directly related to targets



6

Data leakage

Data leakage can happen due to:

Training example leakage (row-wise leakage) – training samples influencing directly testing samples (poor augmentation or data acquisition)

Feature leakage (column-wise leakage) – inclusion of columns directly related to targets

Data acquisition procedures that differ for training data and in-operation data

Cat images were acquired indoors

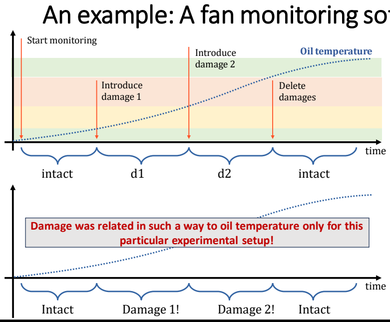
Dog images were acquired outdoors

A classifier recognizing cats from dogs

Is required to recognized new data in all surroundings

7

An example: A fan monitoring software



A lot of features are acquired: vibro features, **oil quality and temperature**, bearings temperatures, electrical features…

The experiment was repeated in this configuration several times…

Then a standard random division into training/validation/testing subsets. Classifier obtains very high accuracy…

Then a CEO is invited for a demonstration…

8

Data diversity

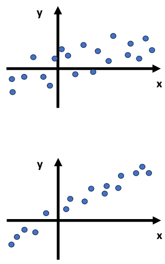
So how to check if our dataset may contain/cause data leakage?

Unfortunately, there are no straightforward methods for that… ☹
But we can:

- Check features and think if they can be measured with no prior knowledge of targets (if no – possible leakage)
- Think about independence of data samples. If they are dependant, testing set should consider a separately acquired batch of data (random split should not be used)
- Check data distribution and think if it is close to one expected in-operation
- Consider order of samples' acquisiton. If they were acquired in particular order, consider using the newest samples for testing purpose only
- Check for experimental conditions and think if they look plausible (if all the expected in-operation conditions are covered)
- Check features and think if some of them would require exactly similar knowledge as targets (e.g. month salary vs year salary)
- If there are different procedures required for acquisition of various classes data, double-check if they influence data features by themselves
- Wherever possible, refrain from random train/test splits in favor of informed splits ensuring independence of samples (mimicking possible future experiment setting)

9

Data visualizations are often not enough



We can often look at two different datasets and roughly estimate the relations in data.

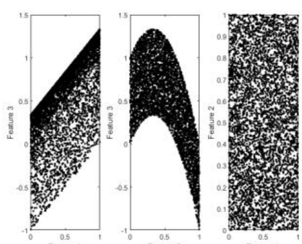
Is it true for high-dimensional datasets?

We actually need a better way of comparing datasets together…

Statistical properties come in mind here: mean values, variances, regression lines, etc.

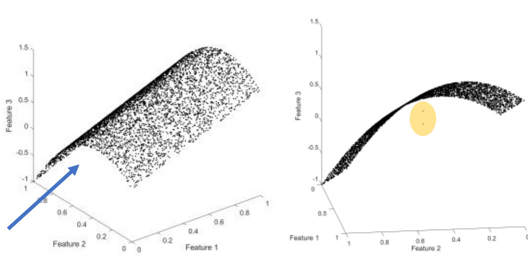
10

Data visualizations are often not enough



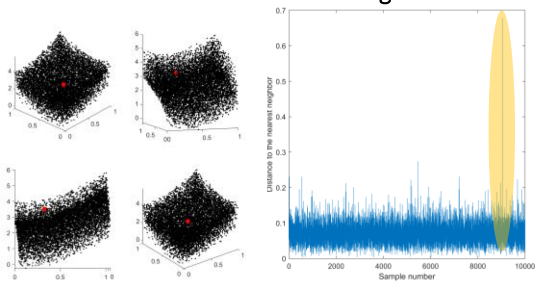
11

Data visualizations are often not enough



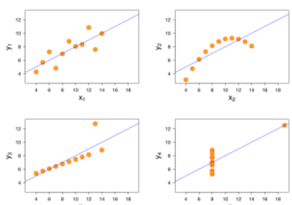
12

Data visualizations are often not enough



13

Statistical measures are often not enough



Anscombe's quartet

While this is a hand-crafted example, it shows the problem behind looking only at statistical data

The same:
- Mean of x
- Variance of x

Almost the same:
- Mean of y
- Variance of y
- Correlation
- Regression line
- Regression error

14

Statistical measures are often not enough

So how to actually look at datasets?

Holistically!

- Look at statistical measures
- Look at raw and processed data
- **Look at outcomes of various models**
- **Draw hypothesis about data and test them**
- Think and experiment!

15

(some) measures of classifiers' efficiency

Consider a classifier for recognizing cats from other animals

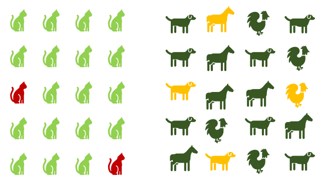
Lets order animals and test our classifier on them.
*We are assuming, that classifier „targets" one class. Say: **cats***
*So it answers a question „Is this a **cat?**". The answer can be **Positive** or **Negative***

*We have 18 **True Positives:***

*4 **False Positives:***

*2 **False Negatives:***

*16 **True Negatives:***



16

(some) measures of classifiers' efficiency

These will form different quality metrics

- ***True Positives:*** Samples correctly assigned a target class label
- ***False Positives:*** Samples incorrectly assigned a target class label (not having a property to which the classifier is tuned)
- ***False Negatives:*** Samples incorrectly assigned as not having target class label (having a property to which the classifier is tuned but missed in classification)
- ***True Negatives:*** Samples correctly identified as not having a target class label

17

(some) measures of classifiers' efficiency

True Positive Rate (TPR):
(Sensitivity, Recall)

$$\frac{TP}{TP+FN}$$

How many cats are we correctly detecting on average out of all cats?
*How **Sensitive** will we be to cats in a cat-crowded room?*
*How well can we **Recall** cat properties to recognize them?*

True Negative rate (TNR):
(Specificity)

$$\frac{TN}{TN+FP}$$

How many other animals can we correctly recognize as not-cats?
*Can we ignore everything that is not a cat knowing cat-**specific** features?*

False Negative Rate:

$$\frac{FN}{FN+TP}$$

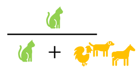
False positive rate (FPR):

$$\frac{FP}{FP+TN}$$

18

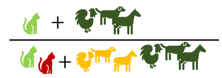
(some) measures of classifiers' efficiency

Precision $\frac{TP}{TP+FP}$



If our classifier detected something as a cat, how likely it is to be correct?
*If it shows us a positive result, is this **Precisely** a cat, not something that is just cat-like?*

Accuracy $\frac{TP+TN}{TP+FP+TN+FN}$



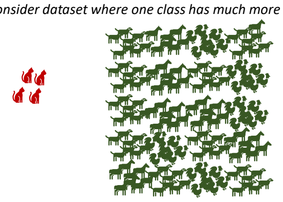
What is a percentage of correct classifications out of all data?

We are introducing just the most popular ones. There are much more metrics out there, look here for example: https://en.wikipedia.org/wiki/Confusion_matrix

19

Why are not using just accuracy?

Consider dataset where one class has much more samples:



Here:

Accuracy is very high
Specificity (TNR) is very high

But…

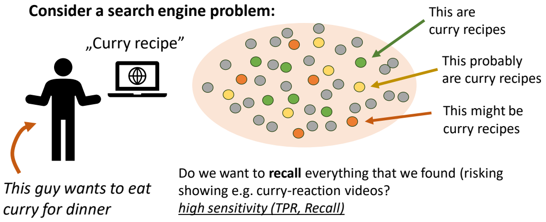
Sensitivity is 0

Our classifier can ALWAYS say „Its not a cat"
and yet the accuracy will be very high…

20

When do we especially need high TPR or high TNR?

Consider a search engine problem:



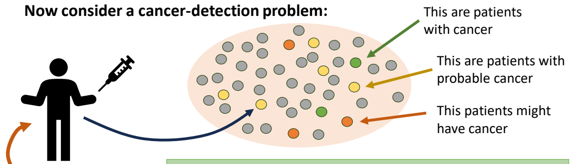
Do we want to **recall** everything that we found (risking showing e.g. curry-reaction videos?
<u>*high sensitivity (TPR, Recall)*</u>

We want to return **precisely** that which is a curry recipe for sure and ignore everything that is not recipe-**specific:**
<u>*high precision, high specificity (TNR)*</u>

21

When do we especially need high TPR or high TNR?

Now consider a cancer-detection problem:



This person is given a test to look for potential cancer symptoms

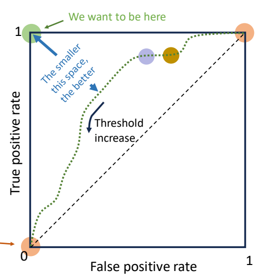
We want to find (and potentially **recall** for further tests) everyone that might potentially have cancer – for further confirmation: high sensitivity (TPR, Recall)

Do we want to treat **precisely** these patients who have very advanced and 100% confirmed cancer and ignore everyone else: high precision, high specificity (TNR)?

22

Receiver Operating Characteristic (ROC)

This is what our classifier does when we change a threshold



If a threshold is low enough, all samples will be classified positively

If a threshold is high enough, all of samples will be classified negatively

We can also use it to pick a reasonable threshold.

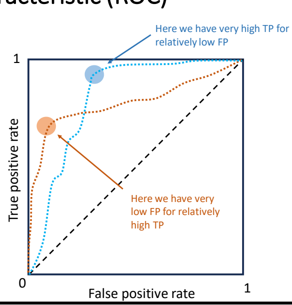
For instance point A is worse than B because at roughly the same TPR we have lover FPR

23

Receiver Operating Characteristic (ROC)

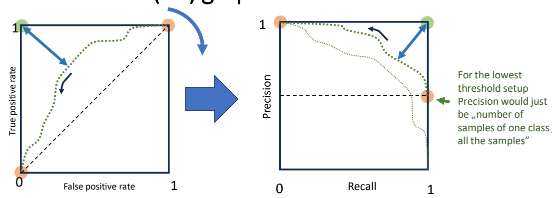
We can also use it to easily compare two classifiers. If low FPR is more important, orange classifier will be better.

If high TPR is more important, blue one will be better



24

Precision-recall (PR) graph



Its not actually rotation! While TPR **is** Recall, Precision **is not** a negative of FPR – but conceptually PR graph serves the same purpose as ROC which is to assess a tradeoff related to focus on increase of TP or decrease of FP

The difference is that ROC focuses on efficiency class-wise (ignoring actual samples' numbers) while PR focuses on actual classifier output. For dissimilar class sizes PR shows what we will actually obtain in practice

25

Precision-recall (PR) graph
Receiver Operating Characteristic (ROC)

How to use them?

1) We can compare classifiers independently from threshold setup (calculation of area-under-curve)

2) We can use these graphs to configure final output with respect to our actual needs

26

What about regression?

Regression is much simpler – you just measure how far your model's prediction is from the actual target:

MAE (Mean absolute error) ← *Easy interpretation (same unit)*

MSE (Mean squared error) ← *Focuses on „important errors"*

RMSE (Root mean squared error) ← *Focuses on „important errors" but is also easy to interpret*

R2 (R Squared, *coefficient of determination*) ← *Allows comparison of models using different datasets*

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Squared errors of your model

Squared errors of baseline model (mean)

27

Confusion matrix

	Actual positive	Actual negative
Predicted positive	TP	FP
Predicted negative	FN	TN

28

Confusion matrix

	Actual A (100)	Actual B (200)	Actual C (100)	Actual D (20)
Predicted A	100	10	30	0
Predicted B	0	150	10	0
Predicted C	0	0	60	0
Predicted D	0	40	0	20

Sum of numbers in column is equal to number of samples in class

Now: if we have B sample, how likely we are to correctly classify it?

We can calculate actual probability of B output given B:

$$P(y_B|B) = \frac{\#\, y_B|B}{\#\, y_A|B + \#\, y_B|B + \#\, y_C|B + \#\, y_D|B,}$$

Number of B outputs for B class (TP)

$$P(y_B|B) = \frac{\#\, y_B|B}{\# B}$$

Number of samples in B (TP + FN)

29

Confusion matrix

We can use confusion matrix for:
- Estimation of probability of correct classification
- Estimation of probability of class presence
- Experiment planning (which classes require more samples)
- Model tuning (which classes require higher accuracy)

Now: classifier says „A" – what does that mean?

If distribution of testing set reflects reality, we can calculate actual probability of A given A output:

$$P(A|y_A) = \frac{\#\, y_A|A}{\#\, y_A|A + \#\, y_A|B + \#\, y_A|C + \#\, y_A|D,}$$

Number of A outputs for A class (TP)

$$P(A|y_A) = \frac{\#\, y_A|A}{\#\, y_A}$$

Number of A outputs (TP + FP)

30

Gentle introduction to bayes rule

Imagine that we have a weather forecast classifier that calculates probability of snow the next day

The classifier is really good, it has overall 95% accuracy,
Snow prediction is as well 95% specific and 95% sensitive
(snow is predicted for 5% of not-snow days and 5% of snow days is not predicted)

The classifier detects snow the next day. What is the probability of this output being correct?

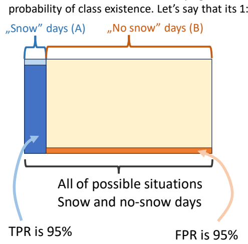
What is the answer if we know that only 10% of days are snowy?

What if we have this answer in the middle of June?

31

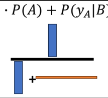
Gentle introduction to bayes rule

We know that there is some underlying probability of class existence. Let's say that its 1:9:



The classifier detects snow the next day. What is the probability of this output being correct?

$$P(A|y_A) = \frac{P(y_A|A) \cdot P(A)}{P(y_A|A) \cdot P(A) + P(y_A|B) \cdot P(B)}$$



In a generalized form this equation is called Bayes conditional probability (Bayes rule):

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

In this equation „A" and „B" refer to general events, not classes in our example! We were denoting B as y_A

32

A hypothesis interpretation to bayes rule

The classifier detects snow the next day. What is the probability of this output being correct?

1. Consider how likely we are to see snow in the first place (10%) — *This is our **prior** probability (assumption before observation)*

2. Consider that we actually have a classifier that tells us „there will be snow" (95% correctness) — *This is new evidence (actual observation). We expect it to UPDATE our prior belief*

3. We combine new evidence with prior assumption to update our belief - we are **more likely** than before to see snow because new evidence tells us so (63%) — *This is our **posterior** probability (updated with new evidence)*

A perfect introduction to graphical interpretation of Bayes rule:
https://youtu.be/HZGCoVF3YvM

33

Things to remember:

1. Data leakage: explain the problem, four main sources for it, provide at least two different practical examples
2. Explain risks related to assessment of data using only visual or only statistical means, explain what are important aspects to consider when looking at a new dataset
3. Define what is a false positive, true positive, false negative and true negative indication, provide a practical example
4. Define classifier metrics: FPR, TPR, FNR, TNR, Precision, Accuracy, Recall, Sensitivity, Specificity
5. Draw examples of ROC and PR diagrams, describe elements and show how threshold setup affects the curves, explain how two different classifiers can be compared on one diagram
6. Explain how ROC and PR diagrams are different from usage perspective and how are they similar (what purpose do they serve)
7. Explain metrics used to measure outcome of regressors
8. Draw an example of a confusion matrix, explain how can it be used to improve experiment or understand classification outcomes
9. Draw a graphical interpretation of a Bayes rule, explain probability of correct classification given TPR, FPR and class prevalence

34