Predictive Modelling - I

Introduction to Classification

Rita P. Ribeiro

Data Mining I - 2023/2024

Summary

- Machine Learning
- Predictive Modelling
  - Classification Problem
  - Regression Problem
- Classification
  - Binary and Multiclass Classification
  - Evaluation Metrics

Machine Learning: Where we at?

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E." - Mitchell, T. (1997)

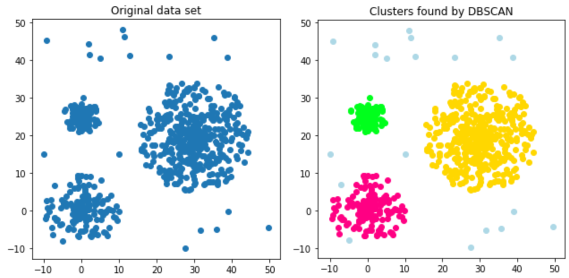
"Machine Learning is the systematic study of algorithms and systems that improve their knowledge or performance with experience"- Flach, P. (2012)

Goal:

- Build models that capture the knowledge from observed cases to make inferences in unseen cases. In principle, more observations should lead to better models!

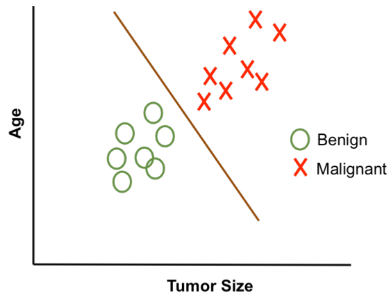
Machine Learning Tasks

- Unsupervised Learning:
  - no target label/value is associated to each example
  - the goal of learning is to obtain a description (e.g. structure, relationships) of the data set
  - e.g. descriptive / predictive clustering, association rules



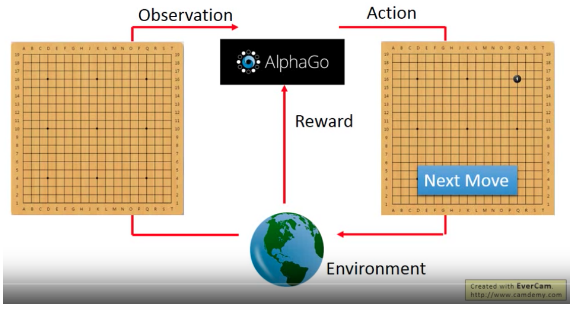
Machine Learning Tasks

- Supervised Learning:
  - there is a target label/value that is associated to each example
  - the goal of the learning task is to learn a function (model) that maps each example with its target variable
  - → Predictive Modelling



Machine Learning Tasks

- Reinforcement Learning:
  - the learning algorithm builds examples from a set of rules; then an iterative process is used to improve (or “reinforce”) the set of examples until some evaluation criterion is good enough.
  - a common example: learn to play chess, make a robot find the best path between two points



Machine Learning Tasks

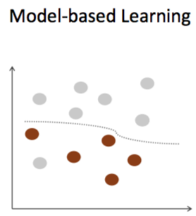
What are the main learning paradigms?

- Batch learning
- Online learning

Is there an assumption on data distribution?

- Parametric
- Non-parametric

What to do when new data points arrive?



Predictive Modelling

Predictive Modelling

Example: Medical Diagnosis}

- Given an historical record containing the symptoms observed in several patients and the respective diagnosis

Headache	Temperature	Age	Throath Inflam.	Flu?
None	36.6	35	No	**No**
None	38.3	40	Yes	**No**
Moderate	38.6	86	No	**Yes**
Strong	40.0	26	No	**Yes**
Strong	38.2	50	No	**Yes**
Strong	36.5	70	Yes	**No**
Moderate	39.1	65	Yes	**Yes**
None	38.3	15	No	**No**
(...)				

- **Predict** the correct diagnosis for a new patient for which we know the symptoms.

Predictive Modelling

- Prediction Models are obtained on the basis of the assumption that there is an unknown mechanism that maps the characteristics of the observations into conclusions/diagnoses.
- The goal of prediction models is to discover this mechanism.
- Medical Diagnosis
  - what we want is to know how symptoms influence the diagnosis
  - use a data set with "examples" of this mapping, e.g. this patient had symptoms s_1, s_2, s_3 and the conclusion was that he had flu.
- Using the available data, obtain a good approximation of the unknown function that maps the observation descriptors into the conclusions

Predictive Modelling

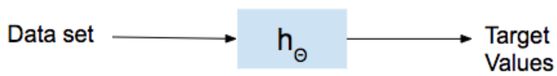
- Descriptors / Predictors / Independent Variables
  - set of variables that describe the properties (features, attributes, predictors) of the cases in the data set
- Target / Dependent Variable:
  - what we want to predict/conclude regards the observations

- It is assumed that the target variable Y is a variable whose values depend on the values of the variables which describe the cases.
- We just do not know how!
- The goal is to obtain an approximation of the function that maps the descriptors to the target variable.

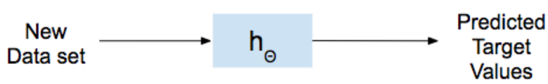
Predictive Modelling

Given a set of predictor variables $\mathbf{X}$ and a target variable Y, there is a function f, such that $f(\mathbf{X}) = Y$

Since f is unknown, the goal is to learn the best approximation to f, h_θ, so that the target values can be obtained from the input data set.

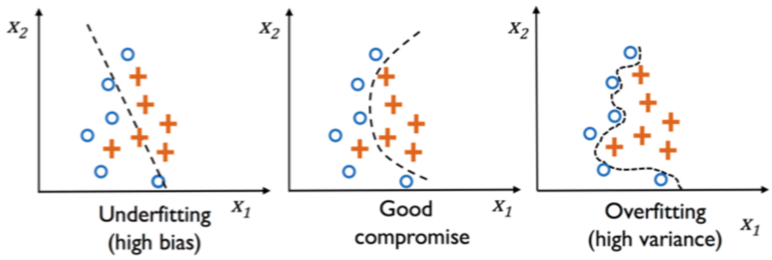


With the built model h_θ, it is possible to make predictions for new, unseen observations!



Predictive Modelling

- Underfitting:
  - model is too simple to capture patterns in data
- Overfitting:
  - model performs well on training data but does not generalize well to unseen data.



Predictive Modelling

Predictive models have two main uses:

1. Prediction:
   - use the obtained models to make predictions regards the target variable of new cases given their descriptors.
2. Comprehensibility:
   - use the models to better understand which are the factors that influence the conclusions.

Predictive Modelling

Types of Prediction Problems

Depending on the type of the target variable Y we may be facing two different types of prediction models:

- Classification Problems
  - the target variable Y is nominal
  - e.g. medical diagnosis - given the symptoms of a patient try to predict the diagnosis
- Regression Problems
  - the target variable Y is numeric
  - e.g. forecast the market value of a certain asset given its characteristics

Prediction Models

- There are many techniques that can be used to obtain prediction models based on a data set.
- Independently of the pros and cons of each alternative, all have some key characteristics:
- They assume a certain functional form for the unknown function $f()$
- Given this assumed form the methods try to obtain the best possible model based on:
  - the given data set
  - a certain preference criterion that allows comparing the different alternative model variants

Prediction Models

- Distance-based approaches
  - e.g. kNN
- Probabilistic approaches
  - e.g. naive Bayes, logistic regression
- Mathematical formulae
  - e.g. linear discriminants, linear regression

Prediction Models

- Logical approaches
  - e.g. classification or regression trees, rules
- Optimization approaches
  - e.g. neural networks, SVMs
- Sets of models (ensembles)
  - e.g. random forests, adaBoost

Prediction Models

These different approaches entail different compromises in terms of:

- assumptions on the unknown form of dependency between the target and the predictors;
- computational complexity of the search problem;
- flexibility in terms of being able to approximate different types of functions;
- interpretability of the resulting model;
- ...

Classification

Classification: Problem Definition

Setting

- $D = \{\langle \mathbf{x_i}, y_i \rangle\}_{i=1}^{N}$
- $\mathbf{x_i} = \langle x_{i1}, x_{i2}, \cdots, x_{ip} \rangle$ the feature vector value
- $y_i \in Y$ is the value of the nominal variable Y

Goal: Learn the best approximation of the unknown function $Y = f(\mathbf{x})$

Approach

- Assume a functional form $h_\theta(\mathbf{x})$ for the unknown function $f()$, where θ are a set of parameters
- Assume a preference criterion over the space θ of possible parameterizations of $h()$
- Search for the "best" $h()$ according to the criterion and the data set

Classification: Problem Definition

Binary Classification Problems (most common)

- when the target variable only assumes two possible values (classes), usually referred as positive and negative class.
- e.g. flu: yes/no, credit: yes/no
- Output of a classification model:
  - class assigned to a case
  - score / probability of case belonging to a certain class; a decision threshold is chosen to establish the predicted class
    - e.g. if $h_\theta(\mathbf{x}_i) \geq 0.5$ then is positive example, otherwise is negative.

Classification: Problem Definition

Multiclass Classification Problems

- when the target variable assumes more than two possible classes
- e.g. insurance risk: low, medium, high
- Some algorithms cannot handle multiclass; the alternative is to combine several binary classifiers
  - one-vs-all: train a model for each class; for k classes, we have k binary classifiers
  - one-vs-one: train a model for each pair of classes; for k classes, we have $k(k-1)/2$ classifiers

Classification: Evaluation Metrics

Goal:

- Obtain reliable estimates of performance and compare different classification models
- Where to assess the performance?
- Which evaluation metrics should be used?

Classification: Evaluation Metrics

Error Rate: proportion of predctions that are incorrect

$$L_{0/1} = \frac{1}{N}\sum_{i=1}^{N} I(\hat{y}_i, y_i)$$

where

- N is the number of cases
- $\hat{y}_i = h_\theta(\mathbf{x_i})$ is the predicted class by the model for the case i
- y_i is the respective true class
- $I()$ is an indicator function such that $I(\hat{y}_i, y_i) = 0$ if $\hat{y}_i = y_i$ and 1 otherwise.

Accuracy = 1 - Error Rate

Classification: Evaluation Metrics

Confusion Matrices

- A square $nc \times nc$ matrix, where nc is the number of class values of the problem
- A special kind of contingency table, with two dimensions (“true class” and “predicted class”)
- Each value reports the number of predictions made by the model of a class for a given true class.

		Predicted Class		
		c_1	c_2	c_3
True Class	c_1	n_{c_1,c_1}	n_{c_1,c_2}	n_{c_1,c_3}
	c_2	n_{c_1,c_1}	n_{c_2,c_2}	n_{c_1,c_1}
	c_3	n_{c_3,c_1}	n_{c_2,c_3}	n_{c_3,c_3}

- The error rate can be calculated from the information on this table.

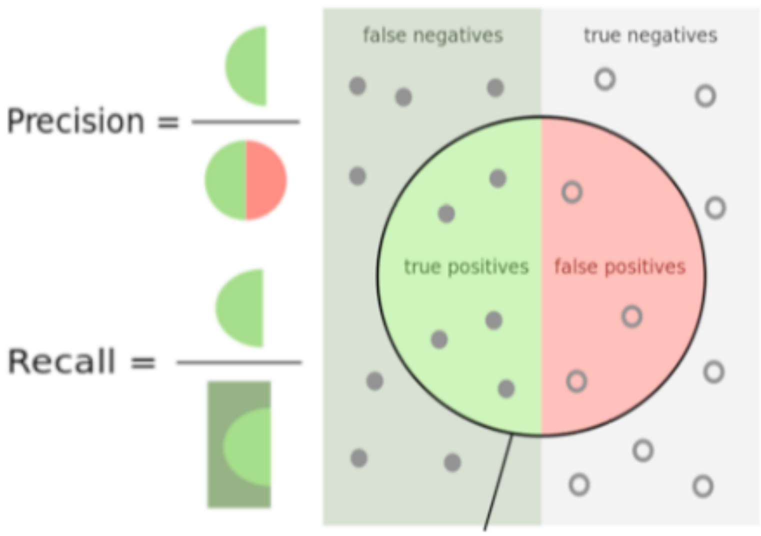
Classification: Evaluation Metrics

Confusion matrix for a binary classification problem

		Predicted Class	
		P	*N*
True Class	*P*	*TP* True Positive	*FN* False Negative
	N	*FP* False Positive	*TN* True Negative

- Accuracy = $\frac{TP+TN}{TP+FP+TN+FN}$
  - proportion of correct predictions
- Precision = $\frac{TP}{TP+FP}$
  - proportion of the positive predictions of the model that are correct
- Recall = $\frac{TP}{TP+FN}$
  - proportion of the positive examples that are captured by the model

Classification: Evaluation Metrics



Adapted from wikipedia

Classification: Evaluation Metrics

- Precision/Recall tradeoff:
  - increasing precision may reduce recall and vice versa.
- It is easy to obtain 100% Recall: always predict P
- F-measure is a statistic that combines Precision an Recall

$$F_\beta = \frac{(\beta^2 + 1) \times Precision \times Recall}{\beta^2 \times Precision + Recall}$$

where β controls the relative importance of Precision and Recall.

- If $\beta = 1$ then is the harmonic mean between Precision and Recall;
- When $\beta \to 0$ the weight of Recall decreases.
- When $\beta \to \infty$ the weight of Precision decreases.

Classification: Evaluation Metrics

- Some classifiers may require higher precision:
  - e.g. classifier that detects videos that are safer for kids. Keep high precision with only safe videos and may reject other videos that are good (low recall).
- Some classifiers may require higher recall:
  - e.g. classifier that detects disease on image samples. High recall to get all disease samples. Can handle some false positives (lower precision) that later will be double checked by doctors.
- There are several tradeoff measures that account for the performance in both classes differently:
  - e.g. G-mean, IBA (Index of Balanced Accuracy)

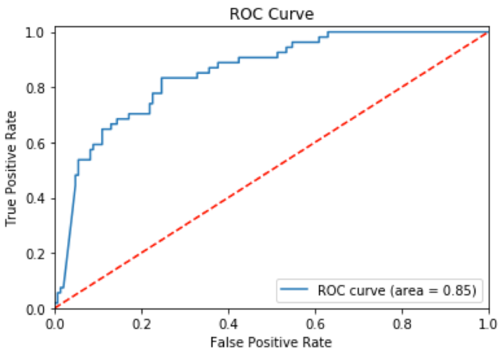
Classification: Evaluation Metrics

- Receiver Operating Characteristic (ROC) Curve:
  - trade-off between *TPR* (*recall*) and *FPR* as the discrimination threshold for the two classes varies.
- False Positive Rate (FPR): $\frac{FP}{TN+FP}$
  - proportion of negative cases wrongly predicted as positive.

True Class	Predicted Probability	FPR	TPR	Thr.
1	0.95			
0	0.92	1/4	1/2	>0.9
0	0.85			
0	0.81	3/4	1/2	>0.8
1	0.78			
0	0.73	4/4	2/2	>0.7

Classification: Evaluation Metrics

- Area Under Curve (AUC) of ROC: performance measure that tells how good the model is in distinguishing the two classes.



- A perfect classifier has AUC of 1;
- A random classifier has AUC of 0.5;
- The higher the AUC, the better.

References

References

Aggarwal, Charu C. 2015. *Data Mining, the Texbook*. Springer.

Ferreira, Pedro G. 2019. "Fundamentals and Applications of ML Course." Slides.

Flach, Peter. 2012. *Machine Learning: The Art and Science of Algorithms That Make Sense of Data*. Cambridge University Press. https://doi.org/10.1017/CBO9780511973000.

Gama, João, André Carlos Ponce de Leon Ferreira de Carvalho, Katti Faceli, Ana Carolina Lorena, and Márcia Oliveira. 2015. *Extração de Conhecimento de Dados: Data Mining -3rd Edition*. Edições Sílabo.

Han, Jiawei, Micheline Kamber, and Jian Pei. 2011. *Data Mining: Concepts and Techniques*. 3rd ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Tan, Pang-Ning, Michael Steinbach, Anuj Karpatne, and Vipin Kumar. 2018. *Introduction to Data Mining*. 2nd ed. Pearson.

Torgo, Luís. 2017. "Data Mining i Course." Slides.