Introduction to Machine Learning

Chapter 2: Machine Learning Tasks

Bernd Bischl, Christoph Molnar, Daniel Schalk, Fabian Scheipl

Department of Statistics – LMU Munich

SUPERVISED LEARNING

- One tries to learn the relationship between “input” x and “output” y.
- For learning, there is training data with labels available
- Mathematically, we face a problem of function approximation: search for an f, such that, for all points in the training data, and also all newly observed points,

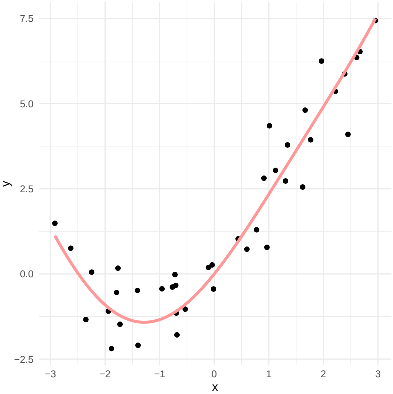
$$y \approx f(x).$$

SUPERVISED LEARNING

Regression Task

Goal: Predict a continuous output

- y is metric variable (with values in $\mathbb{R}$)
- Regression model can be constructed by different methods (e.g. trees or splines), not only statistical (linear) regression!



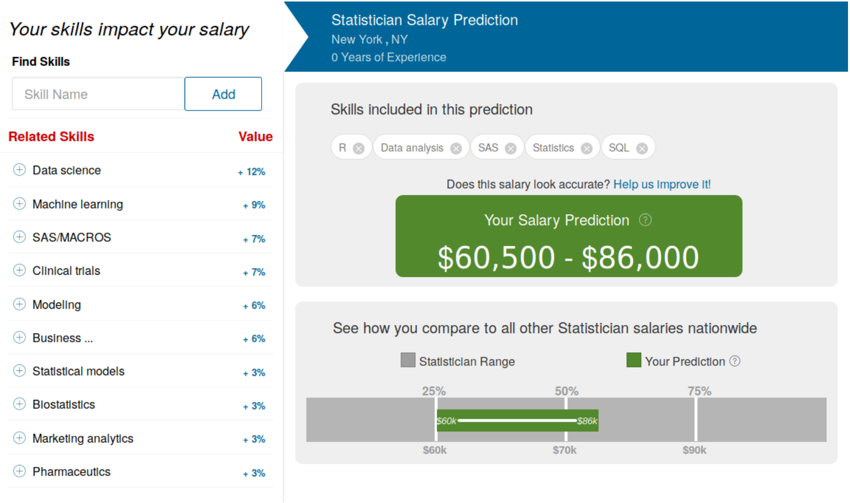
SUPERVISED LEARNING

Regression Task - Examples

- **Stock Trading:** Predicting the exact stock prices on the basis of company data and insider information
- **Pricing:** Anticipating the willingness-to-pay of new customers on the basis of purchases of other customers
- **Medicine:** Calculating the life expectancy for patients with a particular disease and severity (although life time analysis is often better here due to right censoring)
- **Income:** Predicting future income of a person based on education and skills

SUPERVISED LEARNING

Regression Task - Income Prediction



Your skills impact your salary

Find Skills

Skill Name | Add

Related Skills	Value
⊕ Data science	+ 12%
⊕ Machine learning	+ 9%
⊕ SAS/MACROS	+ 7%
⊕ Clinical trials	+ 7%
⊕ Modeling	+ 6%
⊕ Business ...	+ 6%
⊕ Statistical models	+ 3%
⊕ Biostatistics	+ 3%
⊕ Marketing analytics	+ 3%
⊕ Pharmaceutics	+ 3%

Statistician Salary Prediction
New York , NY
0 Years of Experience

Skills included in this prediction

R ⊗ Data analysis ⊗ SAS ⊗ Statistics ⊗ SQL ⊗

Does this salary look accurate? Help us improve it!

Your Salary Prediction ⓘ

$60,500 - $86,000

See how you compare to all other Statistician salaries nationwide

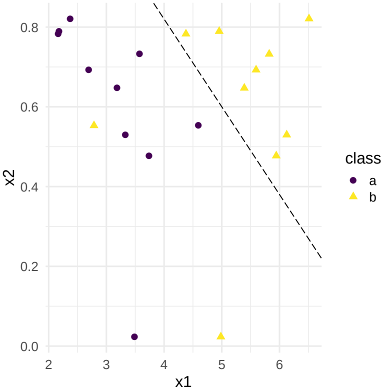
https://www.dice.com/salary-calculator

SUPERVISED LEARNING

Binary Classification Task

Goal: Predict a class (or membership probabilities)

- *y* is a categorical variable with two possible values
- Each observation belongs to exactly one class



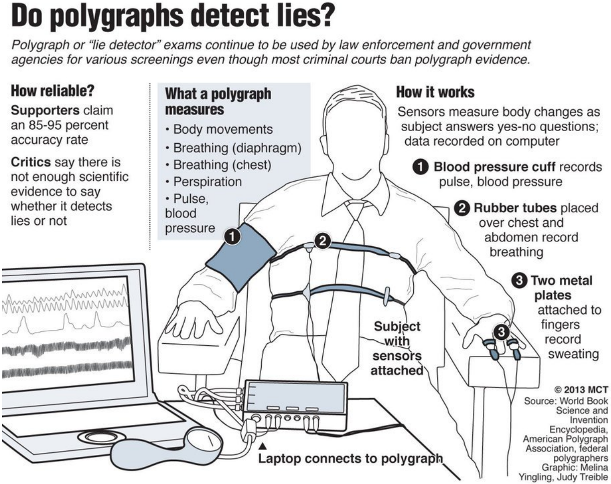
SUPERVISED LEARNING

Binary Classification Task - Examples

- **Credits:** Predicting credit fraud or default risk based on transactions
- **Medical Diagnosis:** Medically testing whether a patient has a specific illness or not
- **Software:** Detecting whether an e-mail is spam or not by using its content
- **Lie Detection:** Determine truthfulness of statements from physiological cues

SUPERVISED LEARNING

Binary Classification Task - Lie Detection



Do polygraphs detect lies?

Polygraph or "lie detector" exams continue to be used by law enforcement and government agencies for various screenings even though most criminal courts ban polygraph evidence.

How reliable?

Supporters claim an 85-95 percent accuracy rate

Critics say there is not enough scientific evidence to say whether it detects lies or not

What a polygraph measures

- Body movements
- Breathing (diaphragm)
- Breathing (chest)
- Perspiration
- Pulse, blood pressure

How it works

Sensors measure body changes as subject answers yes-no questions; data recorded on computer

1. **Blood pressure cuff** records pulse, blood pressure
2. **Rubber tubes** placed over chest and abdomen record breathing
3. **Two metal plates** attached to fingers record sweating

© 2013 MCT
Source: World Book Science and Invention Encyclopedia, American Polygraph Association, federal polygraphers
Graphic: Melina Yingling, Judy Treible

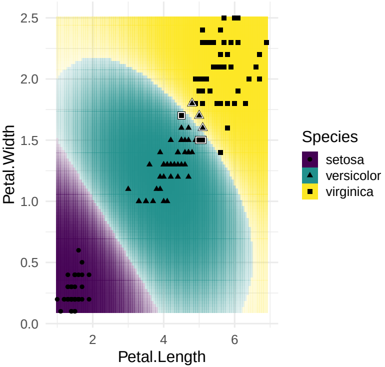
https://www.bendbulletin.com/localstate/deschutescounty/3430324-151/fact-or-fiction-polygraphs-just-an-investigative-tool

SUPERVISED LEARNING

Multiclass Classification Task

Goal: Predict a class (or membership probabilities)

- y is a categorical variable with more than two different unordered discrete values
- Each observation belongs to exactly one class



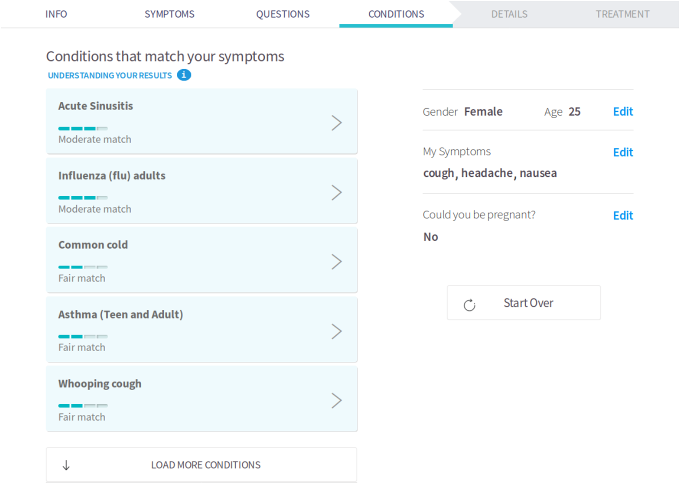
SUPERVISED LEARNING

Multiclass Classification Task - Examples

- **Image Recognition:** Deciding what animal (for example) a picture is showing
- **Stock Trading:** Identifying the best strategy for a specific stock (buy, sell, or wait) based on past prices
- **Biology:** Classifying plants and animals based on their exterior characteristics (e. g. iris flowers)
- **Medical Diagnosis:** Predicting a patients illness using the their symptoms

SUPERVISED LEARNING

Multiclass Classification Task - Medical Diagnosis



Conditions that match your symptoms

Condition	Match
Acute Sinusitis	Moderate match
Influenza (flu) adults	Moderate match
Common cold	Fair match
Asthma (Teen and Adult)	Fair match
Whooping cough	Fair match

Gender Female Age 25

My Symptoms
cough, headache, nausea

Could you be pregnant?
No

https://symptoms.webmd.com

SUPERVISED LEARNING

Classification Models

- Most classification models yield scoring functions for each of the g classes: $f(x) = (f_1(x), \dots, f_g(x)) \in \mathbb{R}^g$.
- These are often called **discriminant functions**, their outputs are class scores or class probabilities.
- The actual classification rule is usually defined as: $h(x) = \underset{k \in \{1,...,g\}}{\arg\max} \ f_k(x)$

SUPERVISED LEARNING

Other supervised learning tasks

- Multilabel classification
- Forecasting
- Survival prediction
- Cost-sensitive classification

ADDITIONAL LEARNING TASKS

Unsupervised learning

- Data without labels y
- Search for patterns within the inputs x
- *unsupervised* as there is no external criterion to optimize or “true” output
- Possible applications:
  - Dimensionality reduction (PCA, Autoencoders ...) : Compress information in $\mathcal{X}$
  - Clustering: Grouping similar observations, separating dissimilar observations
  - Outlier detection
  - Association rules

ADDITIONAL LEARNING TASKS

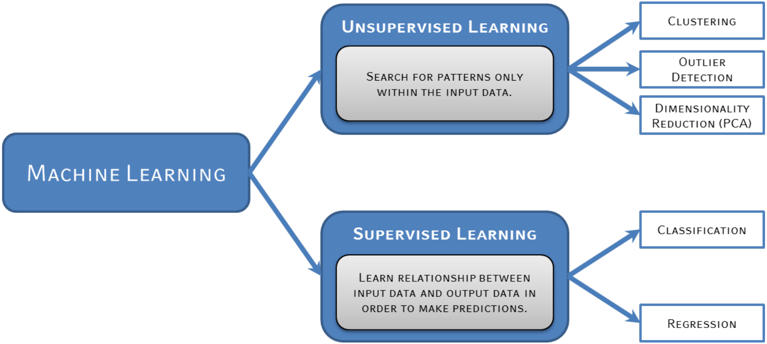
Semi-Supervised learning

- Large amount of labeled data necessary to train reliable model
- Creating labeled datasets often very expensive
- Learn from labeled (expensive) **and** unlabeled (cheap) data
- Unlabeled data in conjunction with a small amount of labeled data improves learning accuracy

Reinforcement learning

- Select actions in subsequent states within a certain environment to maximize lagged future reward
- Example: train neural net to play mario kart (environment)
  - Accelerate/ steer/ break (actions) at each time point (states) during playing
  - Reward: ranking after finish, should be maximized

MACHINE LEARNING TASKS


Machine Learning
Unsupervised Learning
Search for patterns only within the input data.
Clustering
Outlier Detection
Dimensionality Reduction (PCA)
Supervised Learning
Learn relationship between input data and output data in order to make predictions.
Classification
Regression

- In this course, we will deal with **supervised learning** for regression and classification only: predicting y based on x, using a model $f(x)$ that we learned from labeled training data.
- Classification models come with a slight twist: they typically learn g discriminant functions, and then these are turned into discrete predictions (details later).