Machine Learning

4DV660

Welf Löwe

Content

- Overview of the field of artificial intelligence (AI).
- Basic principles for statistical/machine learning (ML).
- Pre-processing of data, feature extraction, dimension reduction.
- Model selection, parameter tuning.
- Regularization of models avoiding over and under-fitting.
- Regression (linear, non-linear)
- Classification (Logistic regression, Nearest neighbor, Naive Bayes, Decision tree, Core methods and support vector machines, Ensemble methods)
- Clustering (k-means, hierarchical)
- Neural networks (reference to the course on Deep Learning later this year).

Goals – You should be able to

- Outline various areas of AI.
- Explain basic principles and applications in ML.
- Describe the weaknesses and advantages of different ML algorithms.
- Describe the different learning paradigms in ML.
- Implement algorithms to solve typical ML problems.
- Represent data to facilitate ML.
- Choose a suitable model for a given problem and evaluate its performance.
- Recognize typical effects of inappropriate initialization values and parameter selection and suggest ways to improve results.
- Recognize cases of over and under-fitting of models and suggest ways to deal with them.
- Reason about effects of, e.g., bias from training data, on actual applications.

Practical Q&A

- Teacher/TA: Welf Löwe, Alisa Lincke, Sebastian Hönel
- Exam:
  - Practical assignment (50%) and an exam (50%)
  - Grades: A, B, C, D, E, Fx, F.
- Literature
  - James, Gareth; Witten, Daniela; Hastie, Trevor; Tibshirani, Robert, An introduction to statistical learning: with applications in R, Springer, 2021, ISBN: 1461471370. 350 pages of 612.
    - Book: https://statlearning.com
    - Data: https://CRAN.R-project.org/package=ISLR (including csv on the MyMoodle the course homepage)
  - Bishop, Christopher, Pattern recognition and machine learning, Springer, 2006, ISBN: 0387310738. 400 pages of 700.
    - Book: https://bit.ly/2RCABxh
  - Compendium with scientific articles. ca 100 pages.
- Course and course material
  - Available in MyMoodle: https://mymoodle.lnu.se/course/view.php?id=57156
  - Mind the updates. So far, the classroom contains material from 2022 (to get an overview of the course)
  - Aligned with Machine learning Lab course (4DV652) and with Deep Learning (4DV661)

Prerequisites

- We assume you have taken an ML/AI course on BSc level
- If not, recap our course for developers
  - Slides: https://coursepress.lnu.se/kurs/applied-machine-learning/
  - Notebooks: https://github.com/WelfLowe/ML4developers

Agenda for today

- What is statistical/machine learning?
  - Why estimate f?
    - Prediction vs inference
    - Regression vs classification
  - How do we estimate f?
    - Parametric vs non-parametric methods
    - Supervised vs unsupervised learning
- Assessing model accuracy
  - Measuring the quality of regression
  - Measuring the quality of classification
- Assignment 1

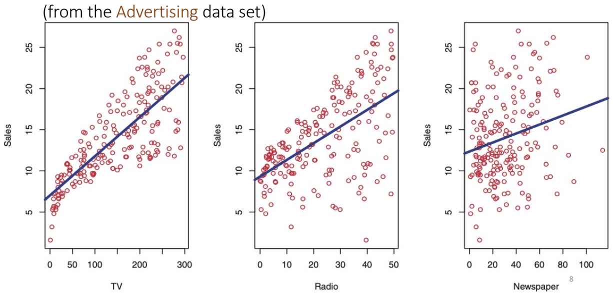
Agenda for today

- What is statistical/machine learning?
  - Why estimate f?
    - Prediction vs inference
    - Regression vs classification
  - How do we estimate f?
    - Parametric vs non-parametric methods
    - Supervised vs unsupervised learning
- Assessing model accuracy
  - Measuring the quality of regression
  - Measuring the quality of classification
- Assignment 1

7

Example: Sales as a function of ad budgets

(from the Advertising data set)



8

In general

- Response (output, label): Y
- Predictors (features, attributes): $X= [X_1, \ldots, X_p]$
- Relationship: $Y = f(X_1, \ldots, X_p) + \varepsilon$
- Irreducible random error: ε
- f represents the systematic information that X provides about Y.

- Statistical/machine learning tries to estimate functions f minimizing the error (i.e., the reducible error, not ε)

9

Reducible and irreducible error

- Approximate $Y = f(X_1, \ldots, X_p) + \varepsilon$ by $\hat{Y} = \hat{f}(X_1, \ldots, X_p)$ where $\hat{f}$ is the estimate, the AI model, for f
- The observed accuracy of a model depends on the reducible (reduced by improved learning approaches and data) and the irreducible (random) error.
- The expectation of the squared error, MSE:

$$\begin{aligned}
E\,(Y - \hat{Y})^2 &= E(f(X) + \varepsilon - \hat{f}(X))^2 \\
&= E(f(X) - \hat{f}(X) + \varepsilon)^2 \\
&= E((f(X) - \hat{f}(X))^2 + 2(f(X) - \hat{f}(X))\varepsilon + \varepsilon^2) \\
&= E(f(X) - \hat{f}(X))^2 + E(2\varepsilon(f(X) - \hat{f}(X))) + E(\varepsilon^2) \\
&= E(f(X) - \hat{f}(X))^2 + 2E(\varepsilon)E(f(X) - \hat{f}(X)) + E(\varepsilon^2) \\
&= E(f(X) - \hat{f}(X))^2 + E(\varepsilon^2) - E(\varepsilon)^2 \\
&= \underbrace{E(f(X) - \hat{f}(X))^2}_{\text{reducible}} + \underbrace{\mathrm{Var}(\varepsilon)}_{\text{irreducible}}
\end{aligned}$$

ε is independent of $f(X) - \hat{f}(X)$

$E(\varepsilon)=E(\varepsilon)^2=0$

10

Why estimate f?

- Prediction:
  - Get a response Y for a given vector of predictors $X= [X_1, \ldots, X_p]$
  - Black box with many predictors, i.e., p large, e.g., deep learning
- Inference:
  - Understand the way the response Y is affected as the predictors $X= [X_1, \ldots, X_p]$ change.
  - White box, e.g., linear models with few predictors, i.e., p small
  - Explainable AI, XAI

11

Prediction

- Predict $Y = f(X_1, \ldots, X_p) + \varepsilon$ by $\hat{Y} = \hat{f}(X_1, \ldots, X_p)$ using the estimator $\hat{f}$
- $\hat{Y}$ is the prediction for Y
- Answers question like:
  - What is the expected value y of Y given the values $x= [x_1, \ldots, x_p]$ for the predictors $X= [X_1, \ldots, X_p]$

12

Advertising Example (cont’d)

- What are the expected sales for a given budget allocated on TV, radio, and newspaper?
- What is the expected increase in sales for an increase in budget allocated on TV, radio, and newspaper?
- …

13

Inference

- Get insights
- Understand the relationship: $Y = f(X_1, \ldots, X_p) + \varepsilon$
- Describe f (and ε)
- Answer questions like:
  - Which predictors are associated with the response?
  - What is the relationship between the response and each predictor?
  - Can the relationship between response and each predictor be adequately summarized using a known type of function, e.g., linear?
- eXplainable AI (XAI)

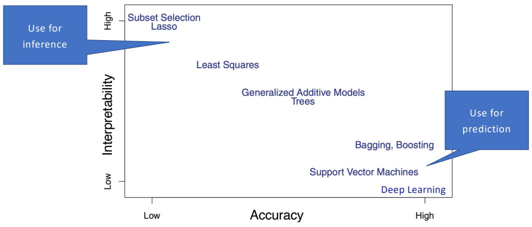
14

Advertising Example (cont’d)

- Which media contribute (most) to sales?
- Which media generate the biggest boost in sales? Is this decision dependent on the total budget?
- Shall we allocate the whole budget on on media or is there a benefit of distributing it to different channels? Is this decision dependent on the total budget?
- …

15

Trade-Off Between Prediction Accuracy and Model Interpretability



16

What do the functions f respond?

- Regression problem:
  - Response Y is quantitative
  - E.g., sales, income, etc.
- Classification problem:
  - Response Y is categorical
  - E.g., a diagnosis (Acute Myelogenous Leukemia, Acute Lymphoblastic Leukemia, or No Leukemia), a person’s gender (male or female), etc.
- Regression problem that is addressed is to be distinguished from the regression method. E.g.:
  - Least squares linear regression (a method) is used to get a quantitative response (kind of problem addressed),
  - Logistic regression (a method) is typically used for a categorial (two-class, binary) response (kind of problem addressed).

17

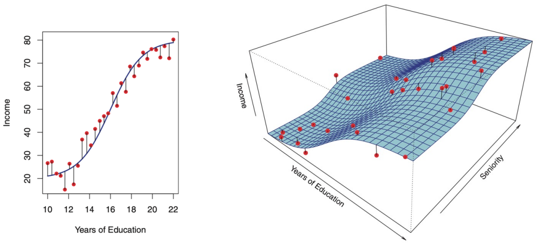
How do we estimate f?

- Parametric methods
  - The function class of f is known
  - The problem of estimating f is reduced to estimating a set of parameters defining a function of this class
  - E.g., linear combination of predictors, $Y = f(X) = A{\cdot}X + b$, estimate A, b
- Non-parametric methods
  - No explicit assumptions about the function class of f
  - Estimate of f based on piece-by-piece functions $f_1 \ldots f_n$ (each with a known function class) as close to the data points as possible without being too “wiggly” (over-fitting)
  - E.g., linear splines, k-means

18

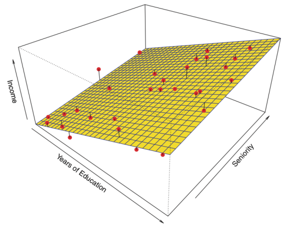
Example: Income as a function education (and seniority)

(from the income data set)



19

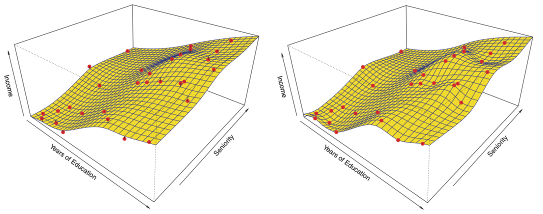
Example (Income): Parametric Model



Linear function fit to the Income data, defined by gradients $A=[a1, a2]$ and offset b.

20

Example (Income): Non-Parametric Model



Smooth (left) and wiggly (right) thin-plate spline fit to the Income data

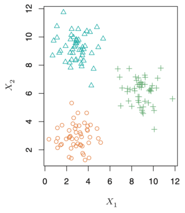
21

How do we estimate f? (cont’d)

- Supervised Learning
  - The response values y^i are known for a set of predictor vector values $x^i= [x^i1, \dots, x^ip]$
  - The pairs of response and predictor values { … (x^i , y^i) … } are called the training data
  - E.g., parametric (linear) and non-parametric (linear-spline-based) regression
- Reinforcement Learning
  - In its simplest form, an online version of supervised learning, where training data is given as a correction of (false) prediction, e.g., in Spam classification
  - In general, only feedback on the prediction accuracy or a corrective direction is given
- Unsupervised Learning
  - The response values are unknown,
  - Only specific problems can be addressed
  - E.g., clustering, pattern extraction, score aggregation

22

Example: clustering as unsupervised learning



Classes (blue, green, orange) are not given
for data points $(X1, X2)$ but
calculated based on properties of the solution, e.g.,
minimize the sum of distances of each datapoint $(x1, x2)$
to its respective cluster mean.

23

Agenda for today

- What is statistical/machine learning?
  - Why estimate f?
    - Prediction vs inference
    - Regression vs classification
  - How do we estimate f?
    - Parametric vs non-parametric methods
    - Supervised vs unsupervised learning
- Assessing model accuracy
  - Measuring the quality of regression
  - Measuring the quality of classification
- Assignment 1

24

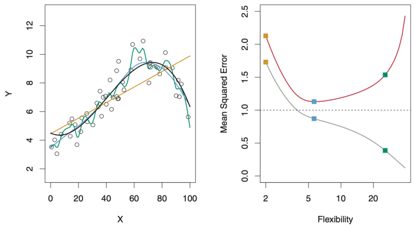
Measuring the quality of regression

- Commonly-used measure is the mean squared error (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

- We are interested in the accuracy of the predictions that we obtain when we apply a trained model to previously unseen test data.
  - Choose the model with the lowest test MSE not the lowest training MSE.
- No guarantee that the method with the lowest training MSE will also have the lowest test MSE
  - still many statistical ML methods specifically estimate coefficients to minimize the training set MSE.

Problem of over-fitting



What to do then?

- Set training data aside (split) to test your estimator
- Cross-validation on different random splits into training and test data

- Visualization before choosing the method and its flexibility (degrees of freedom) if number of features allows
- Set the initial flexibility
- Minimize the test MSE as a function of flexibility, i.e., systematically find the flexibility that minimizes the test MSE
  - E.g., systematically increase the degree of polynomial models: linear, quadratic, cubic, …
- Therefore, we need to understand the Bias-Variance trade-off

The Bias-Variance trade-off

- Variance of $\hat{f}$ refers to the amount by which $\hat{f}$ would change if we estimated it using different training data sets.
- Definition of variance: $\mathrm{Var}(\hat{f}(x_0)) = E(\hat{f}(x_0) - E(\hat{f}(x_0)))^2$
- Bias of $\hat{f}$ is the expected error that is introduced by approximating/simplifying a real-life problem with a model.
- Definition of bias: $\mathrm{Bias}(\hat{f}(x_0)) = E(\hat{f}(x_0)) - E(f(x_0)) = E(\hat{f}(x_0)) - f(x_0)$ as f is deterministic
- For any trainings set (x_0, y_0) the expected test MSE is

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon)$$

Proof

$$\begin{aligned}
E(Y - \hat{Y})^2 &= E(f(X) - \hat{f}(X))^2 + \mathrm{Var}(\varepsilon)\\
&= E(f - \hat{f})^2 + \mathrm{Var}(\varepsilon)\\
&= E(f - \hat{f} - \mathrm{E}(\hat{f}) + \mathrm{E}(\hat{f}))^2 + \mathrm{Var}(\varepsilon)\\
&= E(f - E(\hat{f}) + E(\hat{f}) - \hat{f})^2 + \mathrm{Var}(\varepsilon)\\
&= E((f - E(\hat{f}))^2 + 2(f - E(\hat{f}))(E(\hat{f}) - \hat{f}) + (E(\hat{f}) - \hat{f})^2) + \mathrm{Var}(\varepsilon)\\
&= E(f - E(\hat{f}))^2 + 2E((f - \mathrm{E}(\hat{f}))(E(\hat{f}) - \hat{f})) + E(E(\hat{f}) - \hat{f})^2 + \mathrm{Var}(\varepsilon)\\
&= E(E(\hat{f}) - f)^2 + E(\hat{f} - E(\hat{f}))^2 + 2E((f - E(\hat{f}))(E(\hat{f}) - \hat{f})) + \mathrm{Var}(\varepsilon)\\
&= \mathrm{Bias}(\hat{f})^2 + \mathrm{Var}(\hat{f}) + 2E((f - E(\hat{f}))(E(\hat{f}) - \hat{f})) + \mathrm{Var}(\varepsilon)\\
&= \mathrm{Bias}(\hat{f})^2 + \mathrm{Var}(\hat{f}) + 2(f - E(\hat{f}))E(E(\hat{f}) - \hat{f}) + \mathrm{Var}(\varepsilon)\\
&= \mathrm{Bias}(\hat{f})^2 + \mathrm{Var}(\hat{f}) + 2(f - E(\hat{f}))(E(\hat{f}) - E(\hat{f})) + \mathrm{Var}(\varepsilon)\\
&= \mathrm{Bias}(\hat{f})^2 + \mathrm{Var}(\hat{f}) + 0 + \mathrm{Var}(\varepsilon)
\end{aligned}$$

Rewrite $\hat{f}(X) = \hat{f}$ and $f(X) = f$

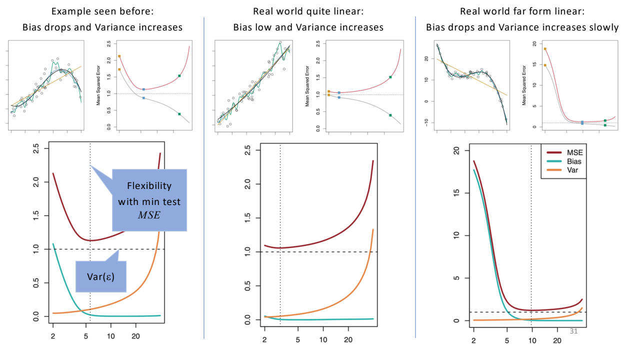
$E(\hat{f}(X)) = \text{const.}$

The Bias-Variance trade-off (cont'd)

- For any value x_0 the expected test MSE is

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon)$$

- Generally, more flexible models result in less bias and vice versa.
- The relative rate of change of variance and bias with flexibility determines whether the test MSE increases or decreases.
  - the squared bias and variance may change at different rates
- The minimum test MSE differs considerably depending on the matching of real-world function with the model and its flexibility.



31

Measuring the quality of classification

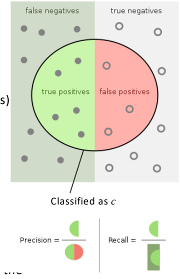
- Corresponding to MSE, in classification the error rate is the proportion of mistakes that are made if we apply our estimate to the training/test observations

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

- $I(cond) = 1$ *if cond, and* 0, *otherwise*
- Min error rate is 0
- As before, choose the model with the lowest test error rate not the lowest training error rate.

32

Precision and Recall



- Precision of class c measures the fraction of garbage (false positives)

$$P_c = \frac{\sum_{i=1}^{n} I(c = y_i = \hat{y}_i)}{\sum_{i=1}^{n} I(c = \hat{y}_i)}$$

- Recall of class c measures the fraction of real classification missed (false negatives)

$$R_c = \frac{\sum_{i=1}^{n} I(c = y_i = \hat{y}_i)}{\sum_{i=1}^{n} I(c = y_i)}$$

- Maximize precision and recall (max$=1$)
  - An increase in precision is often related to a decrease in recall (and vice versa)
- Accuracy of class c is defined as the geometric (harmonic) mean or the weighted geometric mean of precision and recall
  - the most "precise" analysis is not necessarily the most accurate one

33

Why we use Precision/Recall not Error Rate?

- Different costs with misses and garbage classifications in different classes
- Use weighted geometric mean as the measure of quality
- E.g., diagnosis (Leukemia or No Leukemia)
  - In a screening, we could live with some false positive (garbage) Leukemia classifications that are then falsified in detailed exams
  - On the other side we would like to avoid false negative (missing) Leukemia classifications because people die because of a late diagnosis
  - $R_{Leukemia}$ would be weighted higher than $P_{Leukemia}$

34

Agenda for today

- What is statistical/machine learning?
  - Why estimate f?
    - Prediction vs inference
    - Regression vs classification
  - How do we estimate f?
    - Parametric vs non-parametric methods
    - Supervised vs unsupervised learning
- Assessing model accuracy
  - Measuring the quality of regression
  - Measuring the quality of classification
- Assignment 1

35

Rules for the Assignments

- The assignments must be done individually (group work is not allowed)
- You must use the *Moodle assignment submission system* when you hand-in your practical assignments! Email submissions will be ignored.
- The assignments constitute 50% of your final grade. Each assignment is given equal weight (impact factor) on the final grades.
- **Plagiarism**: Each student should hand in an individual set of solutions. Notice, you can exchange ideas between students, not solutions. Any sign of plagiarism will result in an ECTS grade F for all the involved students. As an example, students who copy (parts of the) programs from colleagues or from elsewhere, without giving proper references, fail the course automatically. The same holds for students who let others (friends, relatives, hired skilled persons) complete their assignments for them. Serious cases will be reported to the Disciplinary Board at the Principal's office for further inquires.
- If you don't hand in the assignments before the given deadline your grades will be lowered. Exceptions from this rule can be given if you, within a reasonable time before the actual deadline, contact your TA and give a reasonable explanation.
- Non-approved missed deadlines or an assignment that fails implies that top grade A is no longer possible.
- All the exercises in every assignment are mandatory to complete to receive a passing grade.
- Please contact your teacher/TA if you have any questions regarding these rules.

36

Software for the Assignments

- ML:
  Python3 and appropriate libraries (install them on demand)
  https://www.geeksforgeeks.org/best-python-libraries-for-machine-learning/
- Reporting:
  Jupyter Notebook with Python kernel

Before Assignment 1

Repeat if needed

- Distributions and random variables
- Expected value and variance
- Slides of this Introduction
  - E.g., double-check the proofs on slides 9 and 28

Understand the context:

- Anders Arpteg's AI and Society keynote at Big Data 2020
- Link in myMoodle

Assignment 1:
Setup, data preprocessing and manual introspection

- (Install software)
  - Install Python 3 and write a hello world program.
  - Install Jupyter notebook (standard installation with Python 3 as the kernel language) and write a hello world notebook containing a hello world program.
- Create a Jupyter notebook with text and Python code by following the instructions and the example of "Assignment1-R.pdf" (MyMoodle).
  - **Note** that this PDF document is (an export of) a notebook using "R" as the kernel language. All code snippets must be replaced by Python code.
  - **Note** that the Wage data set that you are supposed to analyze is available for download from the course homepage.
- Deadline: 2023-01-24 (12:00 before the lecture)