Introduction to Machine Learning for NLP I

Benjamin Roth

CIS LMU München

Outline

Course Overview

- Foundations of machine learning
  - loss functions
  - linear regression
  - logistic regression
  - gradient-based optimization
  - neural networks and backpropagation
- Deep learning tools in Python
  - Numpy
  - Theano
  - Keras
  - (some) Tensorflow?, (some) Pytorch?
- Applications
  - Word Embeddings
  - Senitment Analysis
  - Relation extraction
  - (some) Machine Translation?
  - Practical projects (NLP related, to be agreed on during the course)

Lecture Times, Tutorials

- Course homepage:
  `dl-nlp.github.io`
- 9-11 is supposed to be the lecture slot, and 11-12 the tutorial slot ...
- ... but we will not stick to that allocation
- We will sometimes have longer Q&A-style/interactive "tutorial" sessions, sometimes more lectures (see next slide)
- Tutor: Simon Schäfer
  - Will discuss exercise sheets in the tutorials
  - Will help you with the projects

Plan

	9-11 slot	11-12 slot	Ex. sheet
10/18	Overview / ML Intro I	ML Intro I	Linear algebra chapter
10/25	Linear algebra Q&A / ML II	ML II	Probability chapter
11/1	public holiday		
11/8	Probability Q&A / ML III	Numpy	Numpy
11/15	ML IV/Theano Intro	Convolution	Theano I

	9-11 slot	11-12 slot	Ex. sheet
11/22	Embeddings / CNNs & RNNs for NLP	Numpy Q&A	Read LSTM/RNN
11/29	LSTM (reading group)	Theano I Q&A	Theano II
12/6	Keras	Keras	Keras
12/13	DL for Relation Prediction	Theano II Q&A	Relation Prediction
12/20	Word Vectors	Project Topics	Project Assignments

	9-11 slot	11-12 slot	Ex. sheet
1/10	Keras Q&A, Rel.Extr. Q&A	Tensorflow	–
1/17	optimization methods/PyTorch	Help with projects	–
1/24	Other Work at CIS / LMU, Neural MT	Help with projects	–
1/31	Project presentations	presentations	–
2/7	Project presentations	presentations	–

Formalities

- This class is graded by a project
- The grade of the project is determined taking the average of:
  - Grade of the code written for the project.
  - Grade of project documentation / mini-report.
  - Grade of presentation about your project.
  - $\Rightarrow$ You have to pass all three elements in order to pass the course.
- **Bonus points**: The grade can be improved by 0.5 absolute grades through the exercise sheets before New Year.
- Formula:

$$g_{\text{project}} = \frac{g_{\text{project-code}} + g_{\text{project-report}} + g_{\text{project-presentation}}}{3}$$

$$g_{\text{final}} = \text{round}(g_{\text{project}} - 0.5 \cdot \text{x})$$

- where x is the fraction of points reached in the exercises (between 0 and 1), and *round* selects the *closest* value of $1; 1.3; 1.7; 2; \cdots 3.7; 4$

Exercise sheets, Projects, Presentations

- 6 ECTS, 14 weeks
  $\Rightarrow$ avg work load $\sim$ 13hrs / week (3 in class, 10 at home)
  - in the first weeks, spend enough time to read and prepare so that you are not lost later
  - from mid-November to mid-December: programming assignments - coding takes time, and can be frustating (but rewarding)!
- Exercise sheets
  - Work on non-programming exercise sheets individually
  - For exercise sheets that contain programming parts, submit in teams of 2 or 3
- Projects
  - A list of topics will be proposed by me: $\sim$ Implement a deep learning technique applied to information extaction (or other NLP task)
  - Own ideas also possible, need to be discussed with me
  - Work in groups of two or three
  - Project report: 3 pages / team member

Good project code ...

- ... shows that you master the techniques taught in the lectures and exercises.
- ... shows that you can make **"own decisions"**: e.g. adapt model / task / training data etc if necessary.
- ... is well-structured and easy to understand (telling variable names, meaningful modularization – avoid: code duplication, dead code)
- ... is correct (especially: train/dev/test splits, evaluation)
- ... is within the scope of this lecture (time-wise should not exceed $5 \times 10h$)

A good project presentation ...

- ... is short (10 min. p.P. + 15 min. Q&A per team)
- ... similar to the report, contains the problem statement, motivation, model, and results
- ... is targeted to your fellow students, who do not know details beforehand
- ... contains **interesting stuff**: unexpected observations? conclusions / recommendations? did you deviate from some common practice?
- ... demonstrates that all team members worked together on the project
- Possible outline
  - Background / Motivation
  - Formal characterization of techniques used
  - Technical Approach and Difficulties
  - Experiments, Results and Interpretation

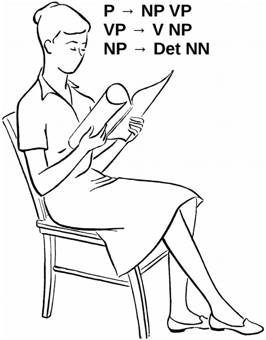
A good project report ...

- ... is concise (3 pages / person) and clear
- ... motivates and describes the model that you have implemented and the results that you have obtained
- ... shows that you can correctly describe the concepts taught in this class
- ... contains **interesting stuff**: unexpected observations? conclusions / recommendations? did you deviate from some common practice?

Outline

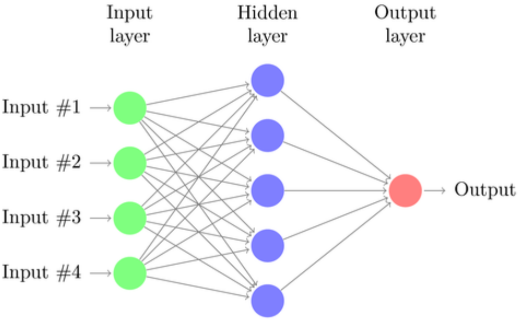
Machine Learning

- Machine learning for natural language processing
  - Why?
  - Advantages and disadvantages to alternatives?
  - Accuracy; Coverage; resources required (data, expertise, human labour); Reliability/Robustness; Explainability



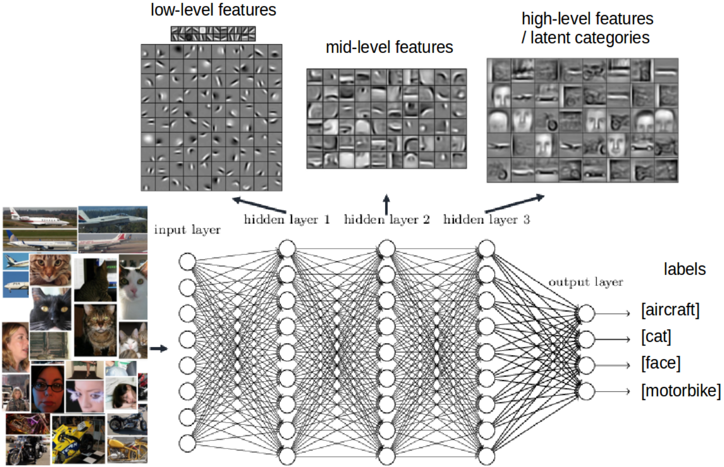
Deep Learning

- Learn complex functions, that are (recursively) composed of simpler functions.
- Many parameters have to be estimated.



Deep Learning

- Main Advantage: Feature learning
  - Models learn to capture *most essential* properties of data (according to some performance measure) as intermediate representations.
  - No need to hand-craft feature extraction algorithms



Neural Networks

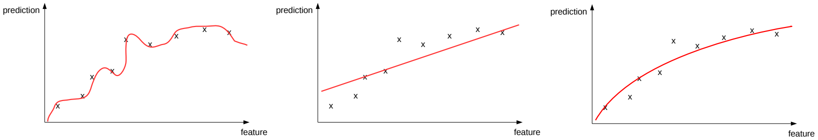
- First training methods for deep nonlinear NNs appeared in the 1960s (Ivakhnenko and others).
- Increasing interest in NN technology (again) since around 5 years ago (*"Neural Network Renaissance"*):
  Orders of magnitude more data and faster computers now.
- Many successes:
  - Image recognition and captioning
  - Speech regonition
  - NLP and Machine translation (demo of Bahdanau / Cho / Bengio system)
  - Game playing (AlphaGO)
  - ...

Machine Learning

- Deep Learning builds on general Machine Learning concepts

$$\text{argmin}_{\boldsymbol{\theta}\in\mathcal{H}} \sum_{i=1}^{m} \mathcal{L}(f(\mathbf{x}_i;\boldsymbol{\theta}), y_i)$$

- Fitting data vs. generalizing from data



Outline

A Definition

"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 1997)

A Definition

"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 1997)

- Learning: Attaining the ability to perform a task.
- A set of examples (*"experience"*) represents a more general task.
- Examples are described by *features*:
  sets of numerical properties that can be represented as vectors $\mathbf{x} \in \mathbb{R}^n$.

Outline

Data

"A computer program is said to learn from **experience** E [...], if its performance [...] improves with **experience** E."

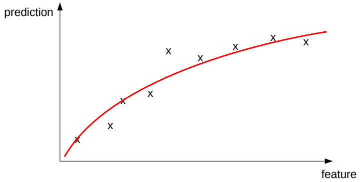
- Dataset: collection of examples
- Design matrix

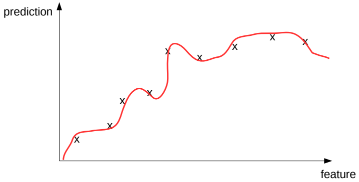
$$\mathbf{X} \in \mathbb{R}^{n \times m}$$

  - n: number of examples
  - m: number of features
  - Example: $X_{i,j}$ count of feature j (e.g. a stem form) in document i.
- Unsupervised learning:
  - Model $\mathbf{X}$, or find interesting properties of $\mathbf{X}$.
  - Training data: only $\mathbf{X}$.
- Supervised learning:
  - Predict *specific* additional properties from $\mathbf{X}$.
  - Training data: Label vector $\mathbf{y} \in \mathbb{R}^n$ together with $\mathbf{X}$

Data

- Low training error does not mean good generalization.
- Algorithm may overfit.





Data Splits

- Best Practice: Split data into training, cross-validation and test set. ("Cross-validation set" = "development set").
  - Optimize low-level parameters (feature weights ...) on training set.
  - Select models and hyper-parameters on cross-validation set. (type of machine learning model, number of features, regularization, priors).
  - It is possible to overfit both in the training as well as in the model selection stage!
  - $\Rightarrow$ Report final score on test set **only after** model has been selected!
- Don't report the error on training or cross-validation set as your model performance!

Outline

Machine Learning Tasks

"A computer program is said to learn [...] with respect to some class of **tasks T** [...] if its performance at **tasks in T** [...] improves [...]"

Types of Tasks:

- Classification
- Regression
- Structured Prediction
- Anomaly Detection
- synthesis and sampling
- Imputation of missing values
- Denoising
- Clustering
- Reinforcement learning
- ...

Machine Learning Tasks: Typical Examples & Examples from Recent NLP Reserch

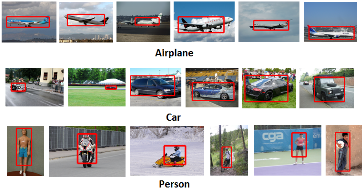
What are the most important conferences relevant to the intersection of ML and NLP?

Task: Classification

- Which of k classes does an example belong to?

$$f : \mathbb{R}^n \to \{1 \dots k\}$$

- Typical example: Categorize image patches
  - Feature vector: color intensities for each pixel; derived features.
  - Output categories: Predefined set of labels



- Typical example: Spam Classification
  - Feature vector: High-dimensional, sparse vector.
    Each dimension indicates occurrence of a particular word, or other email-specific information.
  - Output categories: *"spam"* vs. *'ham"*

Task: Classification

Identifying civilians killed by police with distantly supervised entity-event extraction

Katherine A. Keith, Abram Handler, Michael Pinkham,
Cara Magliozzi, Joshua McDuffie, and **Brendan O'Connor**
College of Information and Computer Sciences
University of Massachusetts Amherst

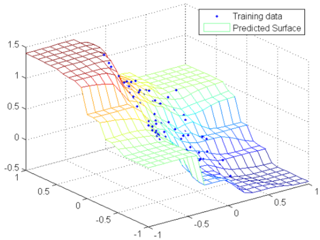
- EMNLP 2017: Given a **person name** in a sentence that contains **keywords** related to **police** ("officer", "police" ...) and to **killing** ("killed", "shot"), was the person a civilian killed by police?

Text	Person killed by police?
Alton Sterling was killed by police.	True
Officers shot and killed **Philando Castile**.	True
Officer **Andrew Hanson** was shot.	False
Police report **Megan Short** was fatally shot in apparent murder-suicide.	False

Task: Regression

- Predict a numerical value given some input.

$$f : \mathbb{R}^n \rightarrow \mathbb{R}$$



- Typical examples:
  - Predict the risk of an insurance customer.
  - Predict the value of a stock.

Task: Regression

Towards an Automatic Turing Test:
Learning to Evaluate Dialogue Responses

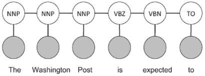
Ryan Lowe[♡*] Michael Noseworthy[♡*] Iulian V. Serban[◇]
Nicolas A.-Gontier[♡] Yoshua Bengio[◇‡] Joelle Pineau[♡‡]

- ACL 2017: Given a response in a multi-turn dialogue, predict the value (on a scale from 1 to 5) how natural a response is.

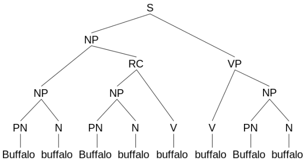
Context	Reference response	Model responses	Human score	ADEM score
photo to see my television debut go to - some. some on <url> - hehe <url> → it really was you? i thought ppl were recognizing someone who looked like you! were the oysters worth the wait?	yeah it was me . haha i 'd kinda forgotten about it it was filmed a while ago	1) i'm not sure. i just don't know what to do with it.	3	1.602
		2) you heard the horsepower productions remix of lee scratch perry's 'exercising' off his 'mighty upsetter' album?	1	1.513
		3) you wont chug a fuzzy peach navel	1	1.744
		4) they were!	5	3.274
just beat call of duty!! → want a cookie? → yes!! → come get it	im in kenmore at the moment	1) i'm gonna get a new phone some moro	1	1.848
		2) no way man.	5	4.265
		3) wow i just got a free pizza coupon! get yours before theres no more! <url>	1	0.921
		4) i'm going to go to the mall.	1	2.634
am i out of twitter jail yet? testing → yeah. i posted bail → thanks. i am a right chatter tweetbox on sundays. same happened last sunday lol	any news on meeting our user ? i go to the us on friday and i don 't want to miss anything arranged	1) i'm not sure if i'm going to be able to get it.	3	1.912
		2) good to see another mac user in the leadership ranks	4	1.417
		3) awww poor baby hope u get to feeling better soon. maybe some many work days at piedmont	2	1.123
		4) did you tweet too much?	5	2.539

Task: Structured Prediction

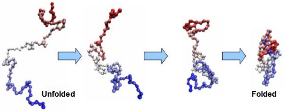
- Predict a multi-valued output with special inter-dependencies and constraints.
- Typical examples:
  - Part-of-speech tagging



  - Syntactic parsing



  - Protein-folding



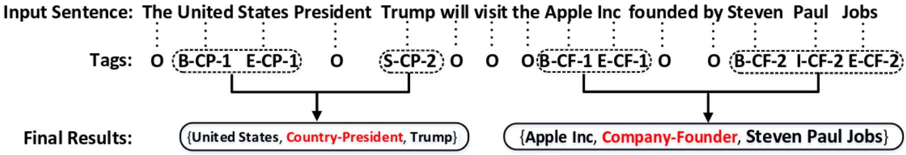
- Often involves search and problem-specific algorithms.

Task: Structured Prediction

Joint Extraction of Entities and Relations Based on a Novel Tagging Scheme

Suncong Zheng, Feng Wang, Hongyun Bao, Yuexing Hao,Peng Zhou, Bo Xu
Institute of Automation, Chinese Academy of Sciences, 100190, Beijing, P.R. China

- ACL 2017: jointly find all relations relations of interest in a sentence by tagging arguments and combining them.



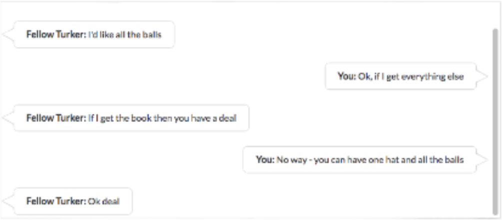
Task: Reinforcement Learning

- In **reinforcement learning**, the model (also called **agent**) needs to select a serious of actions, but only observes the outcome (**reward**) at the end.
- The goal is to predict actions that will maximize the outcome.

Deal or No Deal? End-to-End Learning for Negotiation Dialogues

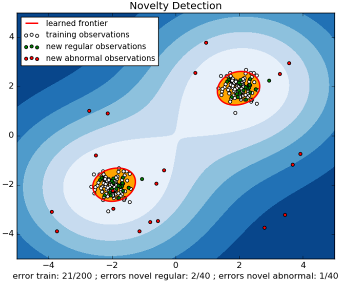
Mike Lewis[1], **Denis Yarats**[1], **Yann N. Dauphin**[1], **Devi Parikh**[2,1] and **Dhruv Batra**[2,1]
[1]Facebook AI Research [2]Georgia Institute of Technology

- EMNLP 2017: The computer negotiates with humans in natural language in order to maximize its points in a game.



Task: Anomaly Detection

- Detect atypical items or events.
- Common approach: Estimate density and identify items that have low probability.



- Examples:
  - Quality assurance
  - Detection of criminal activity
- Often items categorized as outliers are sent to humans for further scrutiny.

Task: Anomaly Detection

Using Automated Metaphor Identification to Aid in Detection and Prediction of First-Episode Schizophrenia

E. Darío Gutiérrez[1] Philip R. Corlett[2] Cheryl M. Corcoran[3] Guillermo A. Cecchi[1]

- ACL 2017: Schizophrenia patients can be detected by their non-standard use of mataphors, and more extreme sentiment expressions.

Supervised and Unsupervised Learning

- Unsupervised learning: Learn interesting properties, such as probability distribution $p(x)$
- Supervised learning: learn mapping from x to y, typically by estimating $p(y|x)$
- Supervised learning in an unsupervised way:

$$p(y|\mathbf{x}) = \frac{p(\mathbf{x}, y)}{\sum_{y'} p(\mathbf{x}, y')}$$

Outline

Performance Measures

"A computer program is said to learn [...] with respect to some [...] **performance measure** P, if its performance [...] **as measured by** P, improves [...]"

- Quantitative measure of algorithm performance.
- Task-specific.

Discrete Loss Functions

- Can be used to measure classification performance.

Discrete Loss Functions

- Can be used to measure classification performance.
- Not applicable to measure density estimation or regression performance.

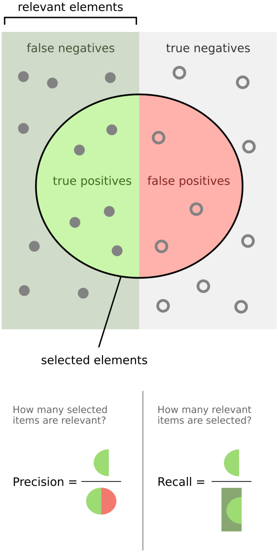
Discrete Loss Functions

- Can be used to measure classification performance.
- Not applicable to measure density estimation or regression performance.
- Accuracy
  - Proportion of examples for which model produces correct output.
  - 0-1 loss = error rate = 1 - accuracy.

Discrete Loss Functions

- Can be used to measure classification performance.
- Not applicable to measure density estimation or regression performance.
- Accuracy
  - Proportion of examples for which model produces correct output.
  - 0-1 loss = error rate = 1 - accuracy.
- Accuracy may be inappropriate for skewed label distributions, where relevant category is rare

$$\text{F1-score} = \frac{2 \cdot \text{Prec} \cdot \text{Rec}}{\text{Prec} + \text{Rec}}$$

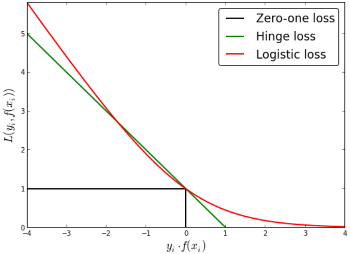


Discrete vs. Continuous Loss Functions

- **Discrete** loss functions cannot indicate **how wrong** a wrong decision for one example is.
- **Continuous** loss functions ...
  - ...are more widely applicable.
  - ...are often easier to optimize (differentiable).
  - ...can also be applied to discrete tasks (classification).
- Sometimes algorithms are optimized using one loss (e.g. Hinge loss) and evaluated using another loss (e.g. F1-Score).

Examples for Continuous Loss Functions

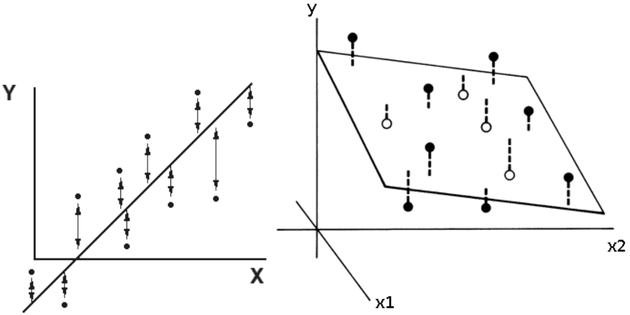
- Density estimation: log probability of example
- Regression: squared error
- Classification: Loss $L(y_i \cdot f(\mathbf{x}_i))$ is function of label$\times$prediction
  - label $\in \{-1, 1\}$, prediction $\in \mathbb{R}$
  - Correct prediction: $y_i \cdot f(\mathbf{x}_i) > 0$
  - Wrong prediction: $y_i \cdot f(\mathbf{x}_i) <= 0$
  - zero-one loss, Hinge-loss, logistic loss ...



- Loss on data set is sum of per-example losses.

Outline

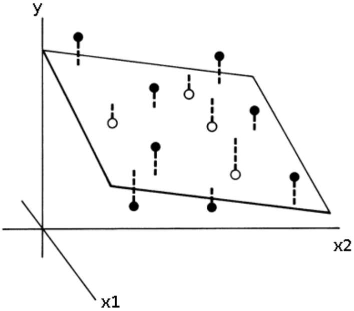
Linear Regression



- For one instance:
  - Input: vector $\mathbf{x} \in \mathbb{R}^n$
  - Output: scalar $y \in \mathbb{R}$
    (actual output: y; predicted output: $\hat{y}$)
  - Linear function

$$\hat{y} = \mathbf{w}^T \mathbf{x} = \sum_{j=1}^{n} w_j x_j$$

Linear Regression



- Linear function:

$$\hat{y} = \mathbf{w}^T \mathbf{x} = \sum_{j=1}^{n} w_j x_j$$

- Parameter vector $\mathbf{w} \in \mathbb{R}^n$
  Weight w_j decides if value of feature x_j increases or decreases prediction $\hat{y}$.

Linear Regression

- For the whole data set:
  - Use matrix **X** and vector **y** to stack instances on top of each other.
  - Typically first column contains all **1** for the intercept (bias, shift) term.

$$\mathbf{X} = \begin{bmatrix} 1 & x_{12} & x_{13} & \dots & x_{1n} \\ 1 & x_{22} & x_{23} & \dots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{m2} & x_{m3} & \dots & x_{mn} \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

- For entire data set, predictions are stacked on top of each other:

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{w}$$

- Estimate parameters using $\mathbf{X}^{(train)}$ and $\mathbf{y}^{(train)}$.
- Make high-level decisions (which features...) using $\mathbf{X}^{(dev)}$ and $\mathbf{y}^{(dev)}$.
- Evaluate resulting model using $\mathbf{X}^{(test)}$ and $\mathbf{y}^{(test)}$.

Simple Example: Housing Prices

- Predict Munich property prices (in 1K Euros) from just one feature: Square meters of property.

$$\mathbf{X} = \begin{bmatrix} 1 & 450 \\ 1 & 900 \\ 1 & 1350 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} 730 \\ 1300 \\ 1700 \end{bmatrix}$$

- Prediction is:

$$\hat{\mathbf{y}} = \begin{bmatrix} w_1 + 450 w_2 \\ w_1 + 900 w_2 \\ w_1 + 1350 w_2 \end{bmatrix} = \begin{bmatrix} 1 & 450 \\ 1 & 900 \\ 1 & 1350 \end{bmatrix} \cdot \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \mathbf{X}\mathbf{w}$$

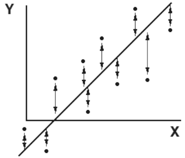
- $\mathbf{w}_1$ will contain costs incurred in any property acquisition
- $\mathbf{w}_2$ will contain remaining average price per square meter.
- Optimal parameters are for the above case:

$$\mathbf{w} = \begin{bmatrix} 273.3 \\ 1.08 \end{bmatrix} \quad \hat{\mathbf{y}} = \begin{bmatrix} 759.1 \\ 1245.1 \\ 1731.1 \end{bmatrix}$$

Linear Regression: Mean Squared Error

- Mean squared error of training (or test) data set is the sum of squared differences between the predictions and labels of all m instances.

$$MSE^{(train)} = \frac{1}{m} \sum_{i=1}^{m} (\hat{y}_i^{(train)} - y_i^{(train)})^2$$



- In matrix notation:

$$MSE^{(train)} = \frac{1}{m} ||\mathbf{\hat{y}}^{(train)} - \mathbf{y}^{(train)})||_2^2$$

$$= \frac{1}{m} ||\mathbf{X}^{(train)}\mathbf{w} - \mathbf{y}^{(train)})||_2^2$$

Outline

Summary

- Deep Learning
  - many successes in recent years
  - feature learning instead of feature engineering
  - builds on general machine learning concepts
- Machine learning definition
  - Data
  - Task
  - Cost function
- Machine tasks
  - Classification
  - Regression
  - ...
- Linear regression
  - Output depends linearly on input
  - Cost function: Mean squared error
- Next up: estimating the parameters