Lecture 3: Linear Regression

Machine Learning, Summer Term 2019

Michael Tangermann   Frank Hutter   Marius Lindauer

University of Freiburg

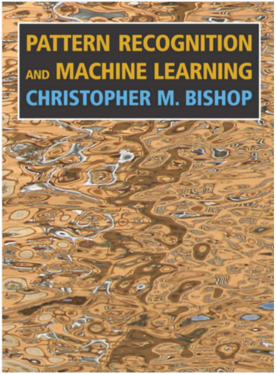
Lecture Overview

1. Brief Recapitulation: Supervised Regression
2. The Linear Regression Model
3. Assumptions and Data Scenarios
4. How to Derive the Parameters...
5. Wrapup: Summary, Related Topics, Preview
6. Let's Work on Assignment 2

Lecture Overview

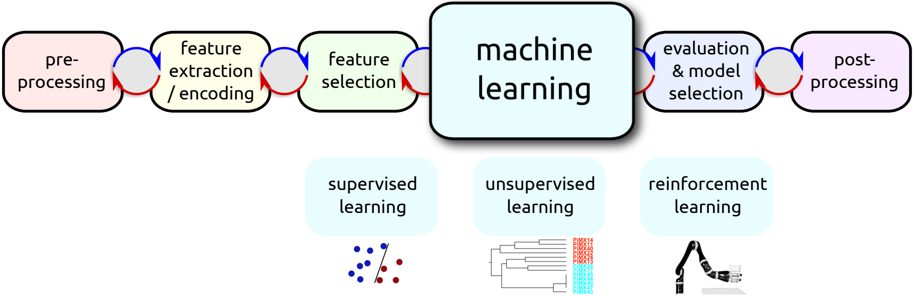
1. Brief Recapitulation: Supervised Regression
2. The Linear Regression Model
3. Assumptions and Data Scenarios
4. How to Derive the Parameters...
5. Wrapup: Summary, Related Topics, Preview
6. Let's Work on Assignment 2

Additional Literature



This book covers linear regression models (today's lecture) and linear discriminant functions (last lecture) nicely, however, it has a dominantly probabilistic approach...

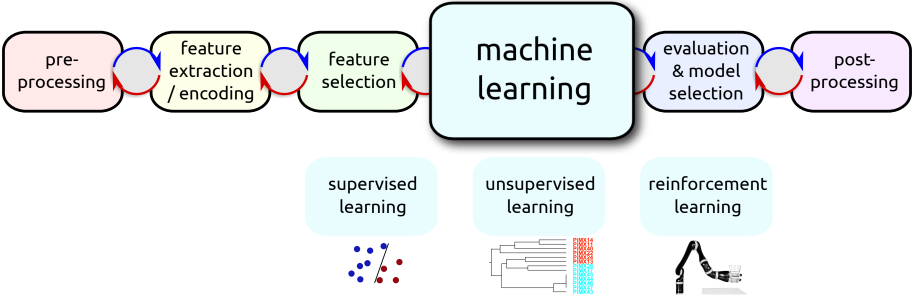
Reminder: ML Design Cycle



Linear regression model are for supervised regression:

- Use past experience to predict the future

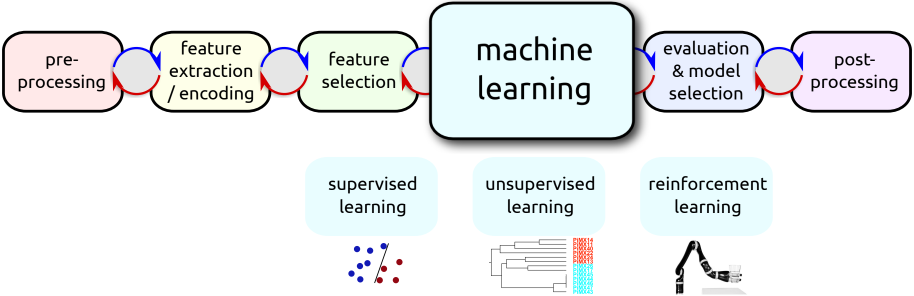
Reminder: ML Design Cycle



Linear regression model are for supervised regression:

- Use past experience to predict the future
- Use labelled data points $\langle(\mathbf{x}_i, y_i)\rangle_{i=1}^{N}$

Reminder: ML Design Cycle



Linear regression model are for supervised regression:

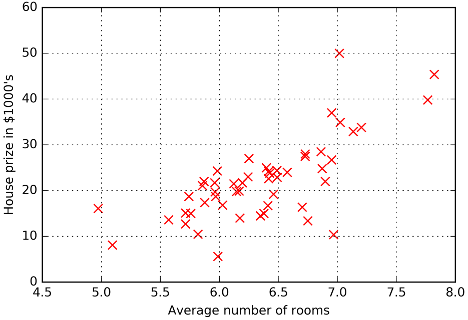
- Use past experience to predict the future
- Use labelled data points $\langle(\mathbf{x}_i, y_i)\rangle_{i=1}^{N}$
- Train a **model** which can predict the label y_{N+1} of a new data point $\mathbf{x}_{N+1}$

Supervised Learning: A Simple Regression Example

Predicting housing prices

- Let's say we only know the average number of rooms in an area
- And we'd like to predict the prize for a house in that area
- One data point: number of rooms $\mathbf{x}_i$ and its prize y_i (in 1000's)

avg. # rooms	y_i
6.575	24
6.377	21.6
5.57	34.7
5.713	33.4
7.024	36.2
5.963	28.7
5.741	22.9
6.417	27.1
...	...

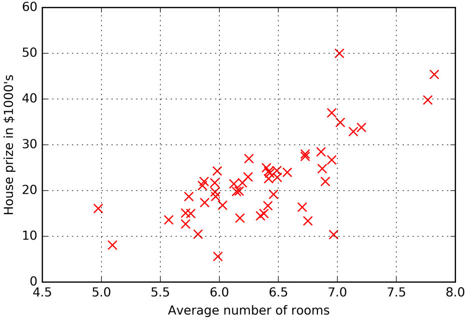


Supervised Learning: A Simple Regression Example

Predicting housing prices

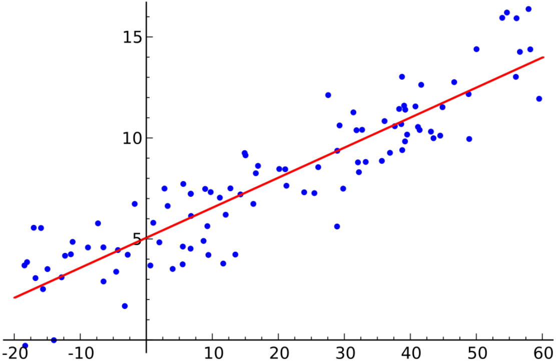
- Let's say we only know the average number of rooms in an area
- And we'd like to predict the prize for a house in that area
- One data point: number of rooms $\mathbf{x}_i$ and its prize y_i (in 1000's)

avg. # rooms	y_i
6.575	24
6.377	21.6
5.57	34.7
5.713	33.4
7.024	36.2
5.963	28.7
5.741	22.9
6.417	27.1
...	...



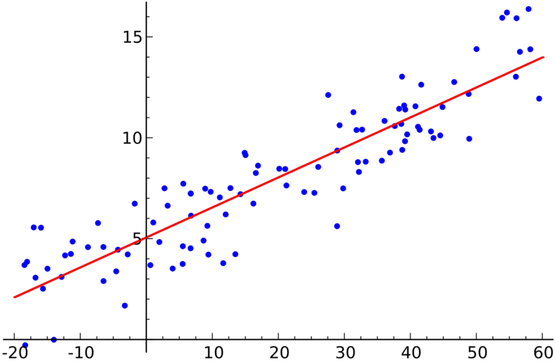
Your ideas for other input variables? 

Reminder: Terminology



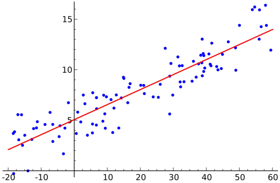
- A data point $\mathbf{x}_i$ is a column vector in $\mathbb{R}^D$
  In the literature, x_i are called "regressors", "covariates" or independent" / "explanatory" / "exogenous" / "explanatory" / "input" / "predictor" variables

Reminder: Terminology



- The label y a scalar value, $y \in \mathbb{R}$
  In the literature, y is called "regressand", or "endogenous" / "response" / "criterion" / "dependent" variable.

Regression Terminology

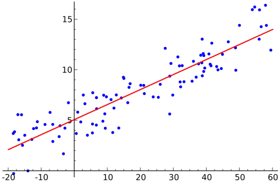


- A data point $\mathbf{x}_i$ is a column vector in $\mathbb{R}^D$
- One label y_i a scalar value, $y_i \in \mathbb{R}$

We are given a labeled data set of N examples:

- $\mathbf{X}$ is a $N \times D$ matrix containing the continuous feature values. ($\mathbf{X}$ contains one transposed data point $\mathbf{x}_i^T$ per row.)
- $\mathbf{y}$ is a $N \times 1$ vector containing the continuous labels.
- Beware: Bishop uses t_i instead of y_i to denominate labels.

Regression Terminology



- In case $\mathbf{x}$ is one-dimensional: *simple* linear regression.
- General case: a data point $\mathbf{x}_i$ is a column vector in $\mathbb{R}^D$
- One label y_i a scalar value, $y_i \in \mathbb{R}$

Remark:

multivariable linear regression or *multiple* linear regression $y_i \in \mathbb{R}$	$\neq$	*multivariate* linear regression $y_i \in \mathbb{R}^D$, with $D > 1$

Regression: The Workflow

- Learn a function $f(\mathbf{x})$, which shall predict the label y based on a data point $\mathbf{x}$ as good as possible.

Regression: The Workflow

- Learn a function $f(\mathbf{x})$, which shall predict the label y based on a data point $\mathbf{x}$ as good as possible.
- Once the function $f(\mathbf{x})$ has been learned, you can infer the continuous label y for a novel input vector $\mathbf{x}$ by simply evaluating the function $f(\mathbf{x})$.

Lecture Overview

Linear Regression: The Basic Idea

The simplest linear model of regression is:

$$f(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + \ldots + w_D x_D$$

where w_0 is a fixed offset, called the *bias*,
$\mathbf{w}$ is the weight vector or parameter vector,
and data point $\mathbf{x} = (x_1, \ldots, x_D)^T$ contains the input variables.

Characteristics:

- Linear function of the weights / parameters w_i
- linear function of the input dimensions / variables x_i
- $\rightarrow$ **significant limitation** of the model!

Linear Regression: The Basic Idea

Which relationsships between $\mathbf{x}$ and y can **not** be described using this simple model? 

(draw here)

Linear Regression with Basis Functions

This limitation can partially be removed by considering linear combinations of fixed nonlinear functions of the input variables:

$$f(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x})$$

where $\phi_j(\mathbf{x})$ are known as *basis functions*.

Characteristics of this *enhanced* linear regression model:

- $f(\mathbf{x}, \mathbf{w})$ is still a linear function of the weights / parameters w_i
- But: $f(\mathbf{x}, \mathbf{w})$ is a **nonlinear** function of the input dimensions / variables x_i

Linear Regression with Basis Functions

This limitation can partially be removed by considering linear combinations of fixed nonlinear functions of the input variables:

$$f(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x})$$

where $\phi_j(\mathbf{x})$ are known as *basis functions*.

Characteristics of this *enhanced* linear regression model (cont.):

- Adding basis functions may enlarge the dimensionality

Linear Regression with Basis Functions

This limitation can partially be removed by considering linear combinations of fixed nonlinear functions of the input variables:

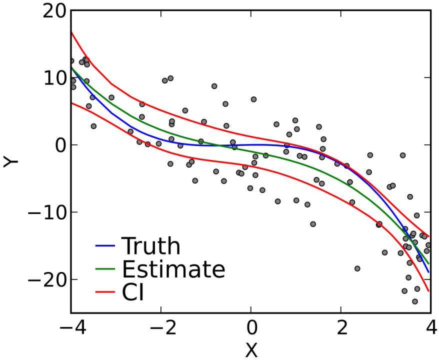
$$f(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x})$$

where $\phi_j(\mathbf{x})$ are known as *basis functions*.

Characteristics of this *enhanced* linear regression model (cont.):

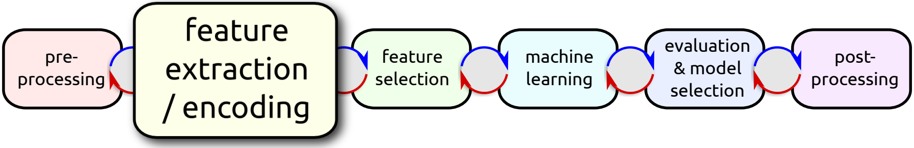
- Adding basis functions may enlarge the dimensionality
- The use of fixed basis functions corresponds to an earlier step in the ML pipeline

Caution with Polynomial Regression / Basis Functions!



- Which risks are involved when using a **polynomial** basis function?

Reminder: ML Design Cycle: Feature Extraction & Encoding



Combine features:

- We can use complex operations to combine features
- Combined features can be more expressive than their components (use expert knowledge!)

See [Bishop, Section 3.1] for common examples on non-linear basis functions.

Linear Regression with Basis Functions

$$f(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x})$$

Please realize

- The index of j is now running up to $M - 1$, thus the number of free parameters is M
- It is often convenient to define an additional dummy basis function $\phi_0(\mathbf{x}) = 1$ to get rid of the bias w_0

Linear Regression with Basis Functions ... and Some Syntactic Sugar

Compare:

$$f(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x})$$

Using one extra basis function $\phi_0(\mathbf{x}) = 1$ delivers a simpler form:

$$f(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x})$$

with $\mathbf{w} = (w_0, \ldots, w_{M-1})^T$ and $\phi = (\phi_0, \ldots, \phi_{M-1})^T$.

In the literature, this is referred to as "augmented notation", and w_0 as the "intercept".

Interpretation of the Weight Parameters $\mathbf{w_i}$

Please discuss: what is the interpretation of a single weight w_i? 

Interpretation of the Weight Parameters $\mathbf{w_i}$

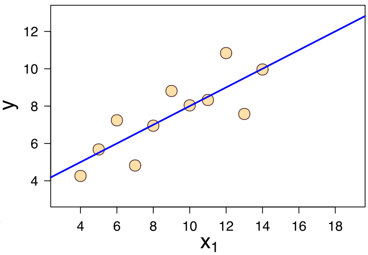
Please discuss: what is the interpretation of a single weight w_i?

Importance / Sensitivity:
Interprete a single weight w_i as a partial derivative with respect to the dependent variable x_i:

- Assume we can keep all other variables fixed – how much would the estimated value for $\hat{y}$ change, if the value of x_1 would be increased/decreased by a value of 1?

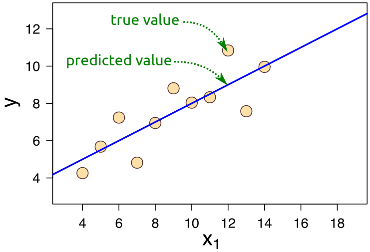
Interpretation of the Quality of the Model

(omitting the basis functions but including augmented $\mathbf{x}$ and $\mathbf{w}$):
How can we guess the quality of the regression model?



Interpretation of the Quality of the Model

(omitting the basis functions but including augmented $\mathbf{x}$ and $\mathbf{w}$):
How can we guess the quality of the regression model?



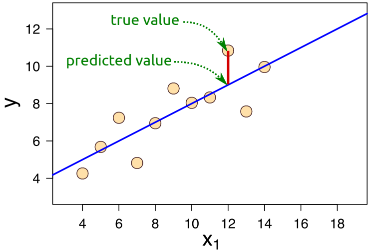
- Error residuals: as the model usually is not perfect, an estimated value $\hat{y}$ may differ from the true y!
- Thus you may find the model expanded to :

$$f(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{D} w_j \mathbf{x}^j + \varepsilon_j = \mathbf{w}^T \mathbf{x} + \boldsymbol{\varepsilon}$$

(with $\mathbf{x}^j$ denoting the j-th dimension of $\mathbf{x}$)

Interpretation of the Quality of the Model

(omitting the basis functions but including augmented $\mathbf{x}$ and $\mathbf{w}$):
How can we guess the quality of the regression model?



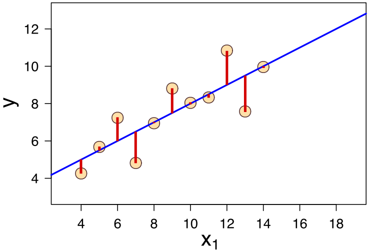
- Error residuals: as the model usually is not perfect, an estimated value $\hat{y}$ may differ from the true y!
- Thus you may find the model expanded to :

$$f(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{D} w_j \mathbf{x}^j + \varepsilon_j = \mathbf{w}^T \mathbf{x} + \boldsymbol{\varepsilon}$$

(with $\mathbf{x}^j$ denoting the j-th dimension of $\mathbf{x}$)

Interpretation of the Quality of the Model

(omitting the basis functions but including augmented $\mathbf{x}$ and $\mathbf{w}$):
How can we guess the quality of the regression model?



- Error residuals: as the model usually is not perfect, an estimated value $\hat{y}$ may differ from the true y!
- Thus you may find the model expanded to :

$$f(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{D} w_j \mathbf{x}^j + \varepsilon_j = \mathbf{w}^T \mathbf{x} + \boldsymbol{\varepsilon}$$

(with $\mathbf{x}^j$ denoting the j-th dimension of $\mathbf{x}$)

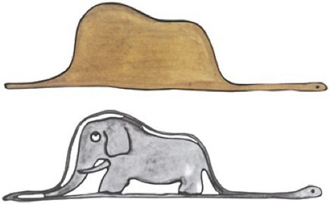
How Can We Derive the Weight Parameters $\mathbf{w_i}$?

Weight vector $\mathbf{w}$ can be derived in (at least) two manners:

How Can We Derive the Weight Parameters $\mathbf{w_i}$?

Weight vector $\mathbf{w}$ can be derived in (at least) two manners:

Option 1: Guess the entries of $\mathbf{w}$ and try to improve them iteratively (gradient descent, other heuristic methods)



Option 2: Determine the weights analytically.

Lecture Overview

Assumptions Required for Analytical Solution

$$y_i = w_0 + w_1 x_1 + \ldots + w_D x_D + \varepsilon_i$$

To formulate the analytical solution for the linear regression model requires to make **three assumptions** about the data and how it can be fitted. If violated, the model may fail or underperform. For $i = 1 \ldots N$ data points:

Assumptions Required for Analytical Solution

$$y_i = w_0 + w_1 x_1 + \ldots + w_D x_D + \varepsilon_i$$

To formulate the analytical solution for the linear regression model requires to make **three assumptions** about the data and how it can be fitted. If violated, the model may fail or underperform. For $i = 1 \ldots N$ data points:

- **A1:** The expected value of the residual errors is zero:
  $\forall i\colon\ \mathrm{E}(\varepsilon_i) = 0$
- **A2:** The residual errors are uncorrelated and share the same variance (across the input range of x):
  $\forall i\colon\ \mathrm{Var}(\varepsilon_i) = \sigma^2$
- **A3:** The residual errors follow a normal distribution:
  $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$

Assumptions Required for Analytical Solution

$$y_i = w_0 + w_1 x_1 + \ldots + w_D x_D + \varepsilon_i$$

To formulate the analytical solution for the linear regression model requires to make **three assumptions** about the data and how it can be fitted. If violated, the model may fail or underperform. For $i = 1 \ldots N$ data points:

- **A1:** The expected value of the residual errors is zero:
  $\forall i: \; \mathrm{E}(\varepsilon_i) = 0$
- **A2:** The residual errors are uncorrelated and share the same variance (across the input range of x):
  $\forall i: \; \mathrm{Var}(\varepsilon_i) = \sigma^2$
- **A3:** The residual errors follow a normal distribution:
  $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$

How would data look like, that violates these assumptions?

(Draw here ...)

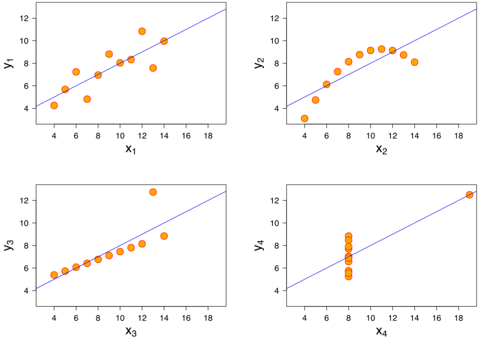
What influences the Regression Function?

- Offsets
- Heteroscedasctic data
- Outliers

Role of outliers

Four datasets generated by the statistician Francis Anscombe in 1973.
They demonstrate

- importance of graphing data before analyzing it
- effect of outliers on statistical properties



[https://en.wikipedia.org/wiki/Anscombe's_quartet]

Lecture Overview

Optimization of $\mathbf{w}$ via Least Squares Loss Function

Assuming the slightly more elegant formulation for linear regression (in augmented vector notation):

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon}$$

we would like to minimize the squared error of the model:

$$\underset{\mathbf{w}}{\operatorname{argmin}} ||\boldsymbol{\varepsilon}||^2 = \underset{\mathbf{w}}{\operatorname{argmin}} ||\mathbf{y} - \mathbf{X}\mathbf{w}||^2$$

Optimization of $\mathbf{w}$ via Least Squares Loss Function

$$
\begin{aligned}
\underset{\mathbf{w}}{\operatorname{argmin}} \; ||\mathbf{y} - \mathbf{X}\mathbf{w}||^2 &= (\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w}) \\
\underset{\mathbf{w}}{\operatorname{argmin}} \; ||\mathbf{y} - \mathbf{X}\mathbf{w}||^2 &= \mathbf{y}^T\mathbf{y} - (\mathbf{X}\mathbf{w})^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\mathbf{w} + (\mathbf{X}\mathbf{w})^T\mathbf{X}\mathbf{w} \\
\underset{\mathbf{w}}{\operatorname{argmin}} \; ||\mathbf{y} - \mathbf{X}\mathbf{w}||^2 &= -(\mathbf{X}\mathbf{w})^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\mathbf{w} + (\mathbf{X}\mathbf{w})^T\mathbf{X}\mathbf{w} \\
&= -\mathbf{w}^T\mathbf{X}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\mathbf{w} + \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} \\
&= -\mathbf{w}^T\mathbf{X}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\mathbf{w} + \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} \\
&= -2\mathbf{y}^T\mathbf{X}\mathbf{w} + \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w}
\end{aligned}
$$

Find optimum by setting the partial derivatives wrt. $\mathbf{w}_i$ to zero.
Use: $\frac{\delta}{\delta \mathbf{w}} \; \mathbf{w}^T A \mathbf{w} = \mathbf{w}^T(A + A^T)$

$$
0 = -2\mathbf{y}^T\mathbf{X} + \mathbf{w}^T(\mathbf{X}^T\mathbf{X} + (\mathbf{X}^T\mathbf{X})^T)
$$

Optimization of $\mathbf{w}$ via Least Squares Loss Function

$$
\begin{aligned}
0 &= -2\mathbf{y}^T\mathbf{X} + \mathbf{w}^T(\mathbf{X}^T\mathbf{X} + (\mathbf{X}^T\mathbf{X})^T) \\
&= -2\mathbf{y}^T\mathbf{X} + 2\mathbf{w}^T\mathbf{X}^T\mathbf{X}
\end{aligned}
$$

$$
\begin{aligned}
\Leftrightarrow \quad \mathbf{y}^T\mathbf{X} &= (\mathbf{w}^T\mathbf{X}^T)\mathbf{X} \\
\Leftrightarrow \quad \mathbf{X}^T\mathbf{y} &= \mathbf{X}^T\mathbf{X}\mathbf{w} \\
\Leftrightarrow \quad (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} &= \mathbf{w}
\end{aligned}
$$

This delivers the analytical solution:

$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{w}$$

- What is the costly part of the model training? 

Optimization of $\mathbf{w}$

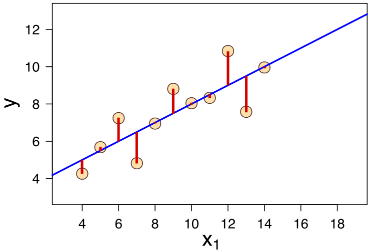
This delivers the analytical solution:

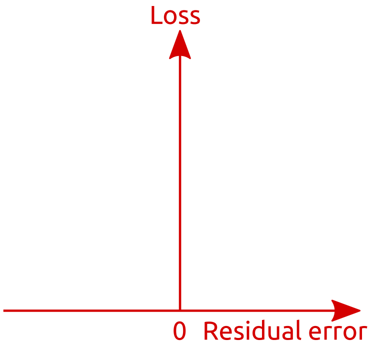
$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{w}$$

- What is the costly part of the model training? 
  $\rightarrow$ Calculation of inverse $\in \mathbb{R}^{D \times D}$
  - Gauss-Jordan elimination procedure: $\mathrm{O}(D^3)$
  - Coppersmith-Winograd: $\mathrm{O}(D^{2.37\dots})$
  - For large dimensions: stochastic gradient descend (nice: the error function is convex!)

Alternative Loss Functions

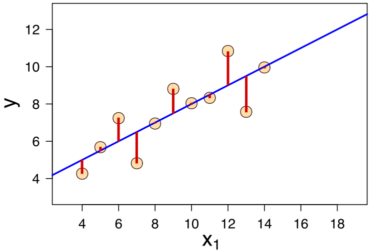
Besides squared loss (L_2), other loss functions are possible. They have different pros and cons.

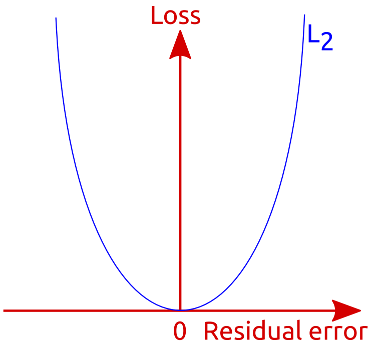




Alternative Loss Functions

Besides squared loss (L_2), other loss functions are possible. They have different pros and cons.

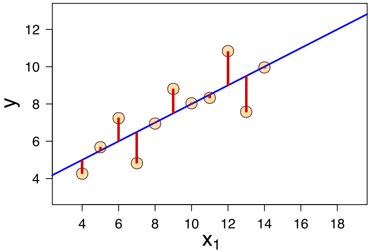


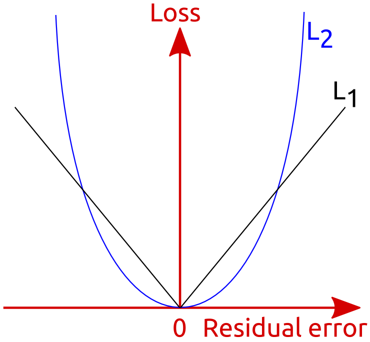


- In which situations would you expect absolute loss (L_1) to be preferable compared to squared loss (L_2)?

Alternative Loss Functions

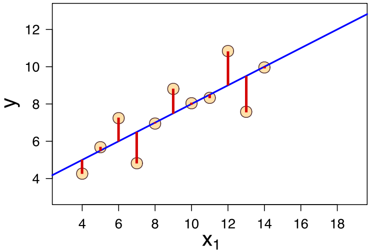
Besides squared loss (L_2), other loss functions are possible. They have different pros and cons.

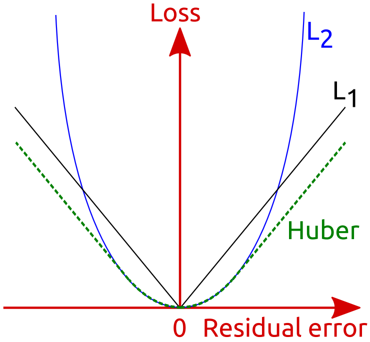




Alternative Loss Functions

Besides squared loss (L_2), other loss functions are possible. They have different pros and cons.





Regularization with Penalty Terms

Loss functions can be combined with penalties for large weight vectors $\mathbf{w}$. So-called penality terms are utilized to regularize the optimization problems.
(Regularization may limit overfitting!). Famous regression models:

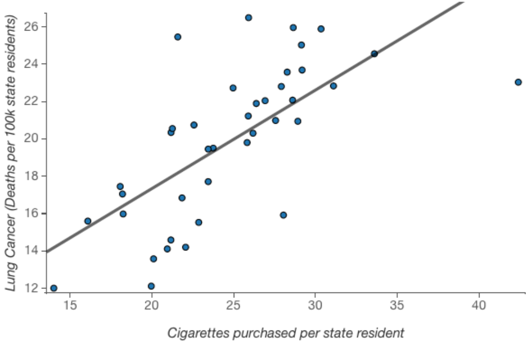
- **Ridge regression**: quadratic loss on residuals with L_2 norm penalty on the weights. Analytic formulation for $\mathbf{w}$. Strong influence of outliers.

- **Lasso**: quadratic loss on residuals with L_1 norm penalty on the weights. No analytic solution, but sparse in $\mathbf{w}$. Reduced influence by outliers.

For most-used combinations, their pros and cons, see e.g.

[http://www.cs.cornell.edu/courses/cs4780/2015fa/web/lecturenotes/lecturenote10.html]

Typical Use of Linear Regression



- To predict unknown values y_i for novel input variables $\mathbf{x}_i$
- Estimate the influence of a single input variable or several variables (e.g. in medicine), i.e. estimating the strength of the **correlation** between x_i and y. **BUT: does not allow for causal interpretation!**
- Visualization to understand a novel dataset: linear or non-linear relationships? Outliers? Distribution of residual errors?

Lecture Overview

Summary by learning goals

Having heard this lecture, you can now . . .

- describe a regression problem
- explain pros and cons of using non-linear basis functions
- formulate the regression function (in augmented and non-augmented form)
- explain the meaning / interpretation of the weight vector
- explain (different) assumptions for linear regression models and effects that violations may have
- formulate the optimization criterion for ordinary least-square regression
- describe pros and cons of different loss functions and regularizations
- derive a regression model from given data and apply it to new data

Organizational Issues

Exam-related:

- Date of the exam: **August 22nd, 2pm**

Organizational Issues

Exam-related:

- Date of the exam: **August 22nd, 2pm**
- The overview sheets for the different algorithms are meant to help you to prepare the exam, but will not be allowed to bring into the exam.

Organizational Issues

Exam-related:

- Date of the exam: **August 22nd, 2pm**
- The overview sheets for the different algorithms are meant to help you to prepare the exam, but will not be allowed to bring into the exam.

Assigment-related:

-  Let's discuss: why do we offer assignments at all, if they are not obligatory?

Lecture Overview

"..."