Statistical Machine Learning

Lecture 2 – Linear regression, regularization

UPPSALA
UNIVERSITET

Thomas Schön
Division of Systems and Control
Department of Information Technology
Uppsala University.

Email: thomas.schon@it.uu.se

Summary of Lecture 1 (I/III)

What is this course about? Supervised machine learning

In one sentence:

Methods for automatically learning (training, estimating, ...)
a model for the relationship between

- the **input** $\mathbf{x}$, and
- the **output** y

from observed **training data**

$$\mathcal{T} := \{(y_1, x_1), (y_2, \mathbf{x}_2), \ldots, (y_n, \mathbf{x}_n)\}.$$

Summary of Lecture 1 (II/III)

Regression vs. classification

- **Quantitative** variables take on numerical values (real numbers, integer values, . . .).
- **Qualitative** variables take on values in one of K distinct classes, e.g. “true or false”, “disease type A, B or C”.

Regression is when the output y is quantitative

Classification is when the output y is qualitative

Summary of Lecture 1 (III/III)

1. Introduction
2. **Linear regression, regularization**
- Introduction to Python & scikit-learn
3. Classification, logistic regression
4. Classification, LDA, QDA, k-NN
5. Bias-variance trade-off, cross validation
6. Tree-based methods, bagging
7. Boosting
8. Deep learning I
9. Deep learning II
10. Summary and guest lecture

”Warm-up videos” for each lecture linked from the home page.
Watch before coming to the lecture (the evening before, or so).

UPPSALA UNIVERSITET

Outline – Lecture 2

Aim: To introduce linear regression and its regularized version.

Outline:

1. Summary of Lecture 1
2. Linear regression models
3. Maximum likelihood and least squares
4. Regularization
   - Ridge regression
   - LASSO

Linear regression is a "working horse" of statistics and (supervised) machine learning.

Qualitative inputs

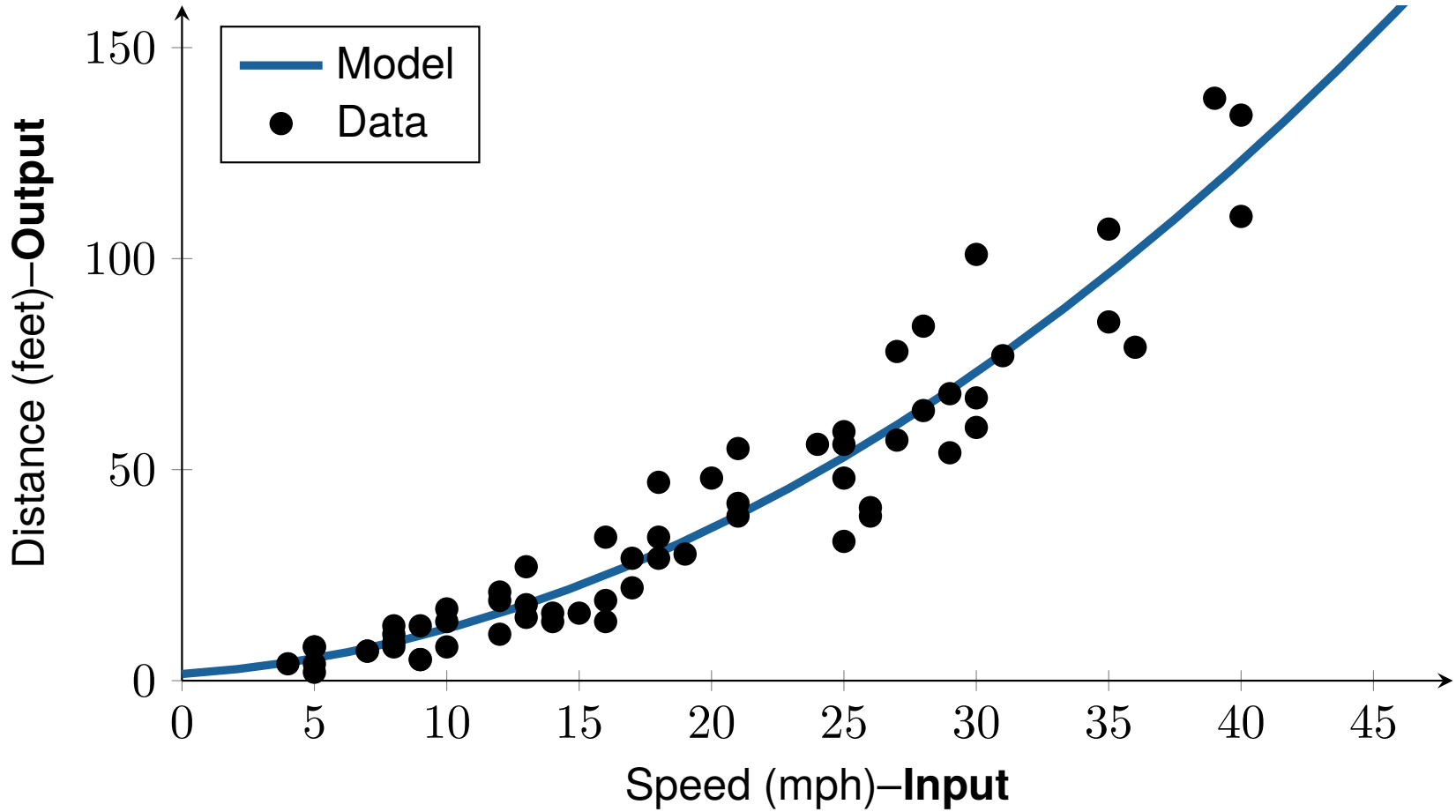
Qualitative or quantitative?

17.31 kg, 22.37 kg, 51.34 kg	Quantitative
1 = brown hair, 2 = red hair, 3 = blonde hair	Qualitative
Adenine, Thymine, Cytosine, Guanine	Qualitative
1 bike, 2 bikes, 5 bikes	Quantitative

Qualitative output variable? $\rightarrow$ *classification*

Qualitative input variable? Still regression!

Regression example: car stopping distances



(in fact a linear regression model with nonlinear transformation of the input variables)

Regression example: Alpha Go zero

- Input: State of the game (19×19 grid, either black, white or blank)
- Output: Probability for the current player to win the game
- \+ *reinforcement learning*

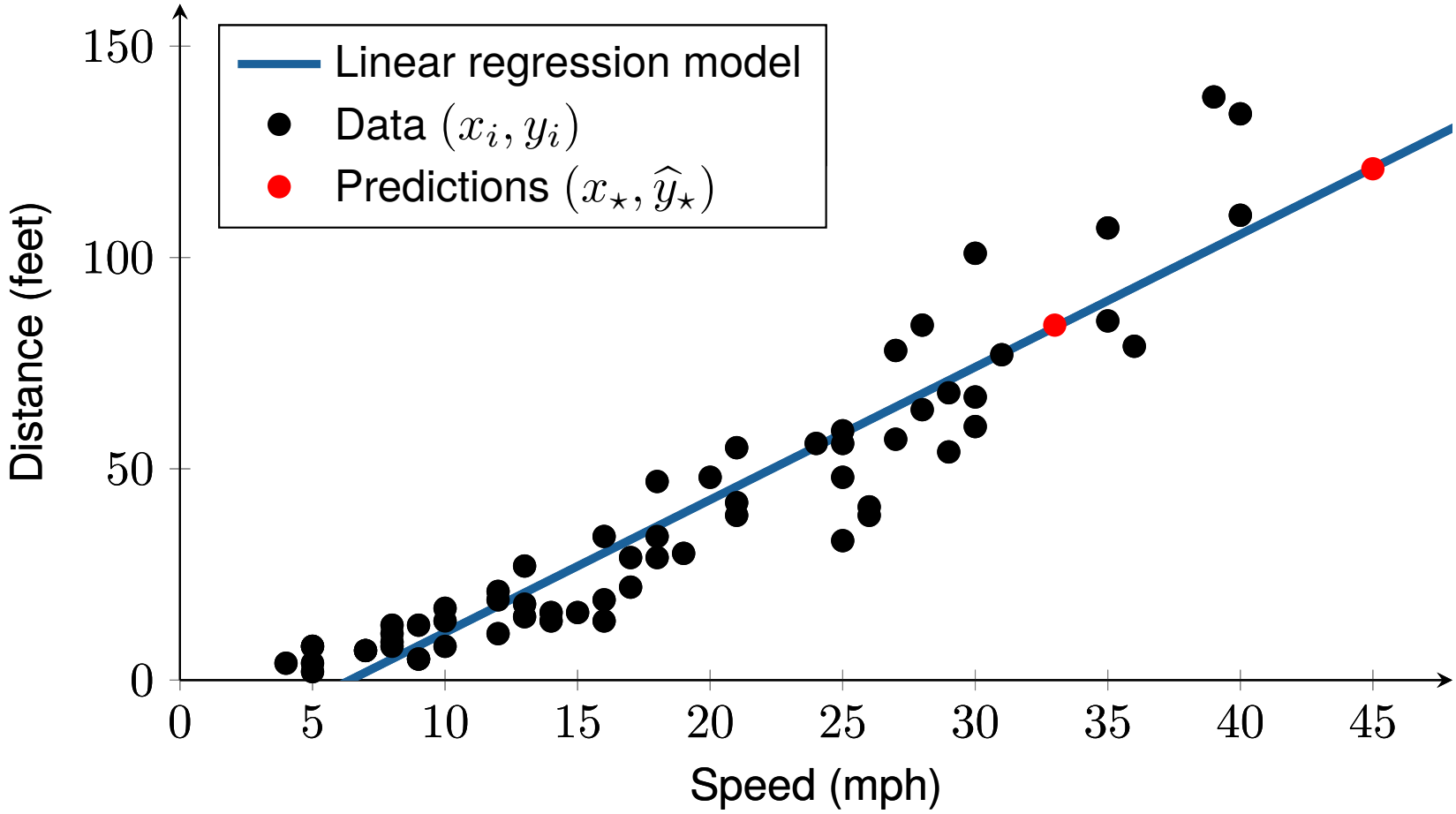
Silver et al. **Mastering the game of Go with deep neural networks and tree search**, *Nature* 529, 484–489, 2016.

- Input: Same
- Output: Probability for the current player to win the game *and* what move to make

Silver et al. **Mastering the game of Go without human knowledge**, *Nature* 550, 354–359, 2017.

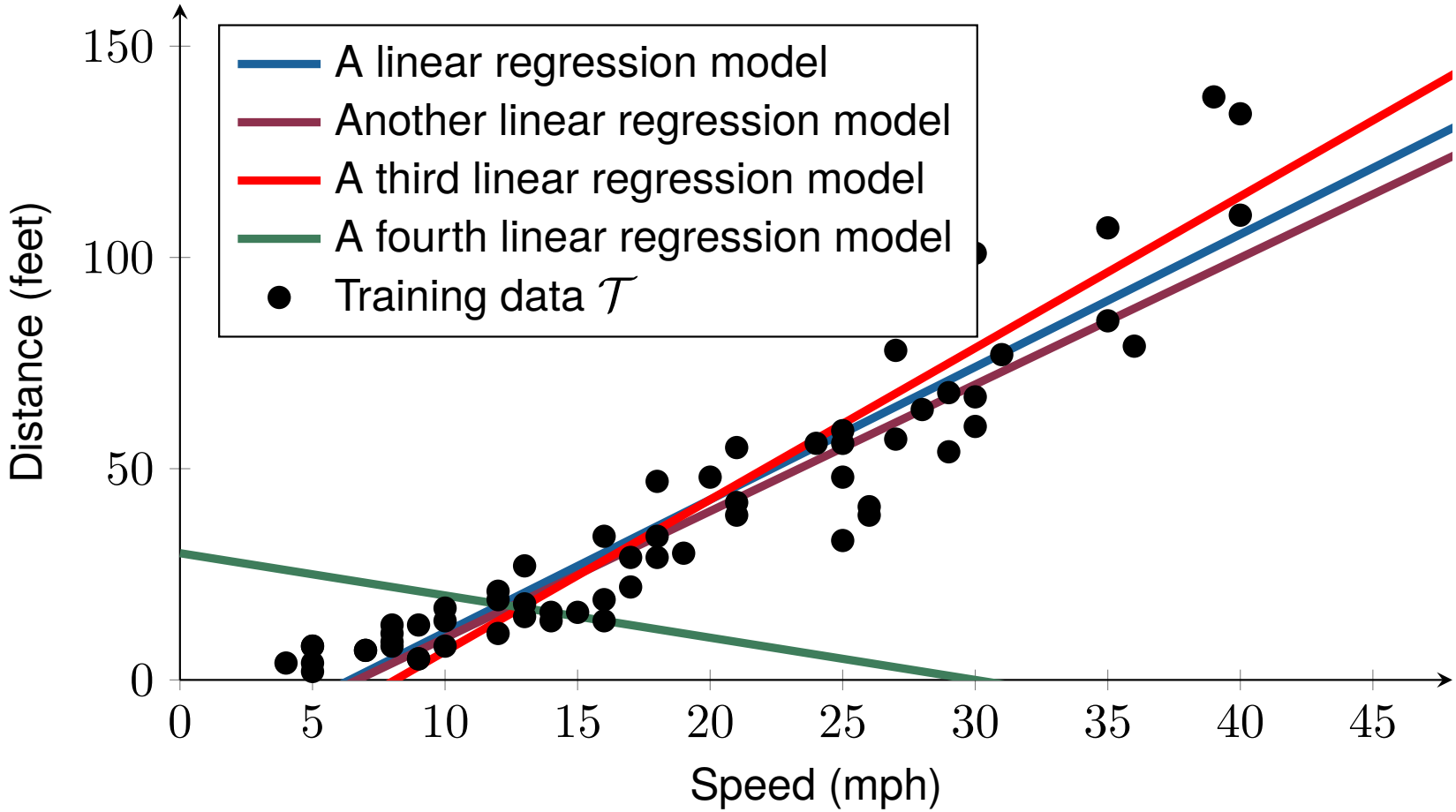
Silver et al. **A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play**, *Science*, 362(6419): 1140–1144, 2018.

Linear regression ($p = 1$)



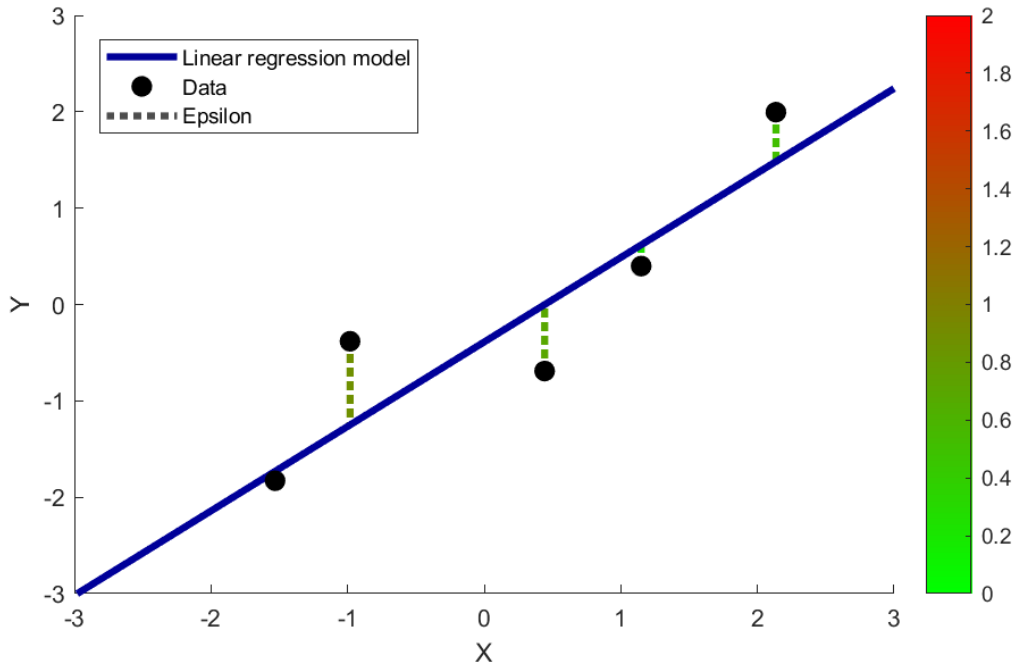
UPPSALA UNIVERSITET

What is a good model?



Learning using maximum likelihood

Learning a model from data is a matter of looking at the errors ε!



Maximum likelihood: Think of ε (dotted) as random variables, and ***choose the model*** (solid) ***such that the resulting*** ε ***are as likely as possible***.

Linear regression model again

Recall our linear regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma_\varepsilon^2 I).$$

Assumptions (for now):

1. $\mathbf{y}$ – observed **random** variable.
2. $\boldsymbol{\beta}$ – unknown **deterministic** variable.
3. $\boldsymbol{X}$ – known **deterministic** variable.
4. $\boldsymbol{\varepsilon}$ – unknown **random** variable.
5. σ_ε – unknown/known **deterministic** variable.

Learning using maximum likelihood

The least squares problem

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}} \|\mathbf{X}\beta - \mathbf{y}\|_2^2$$

is solved by the **normal equations**

$$\mathbf{X}^\mathsf{T}\mathbf{X}\widehat{\beta} = \mathbf{X}^\mathsf{T}\mathbf{y}.$$

Linear regression: the key concepts

The linear regression model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$$

+

Maximum likelihood

$$\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2) \text{ iid}$$

} **Our first learning tool**

Example

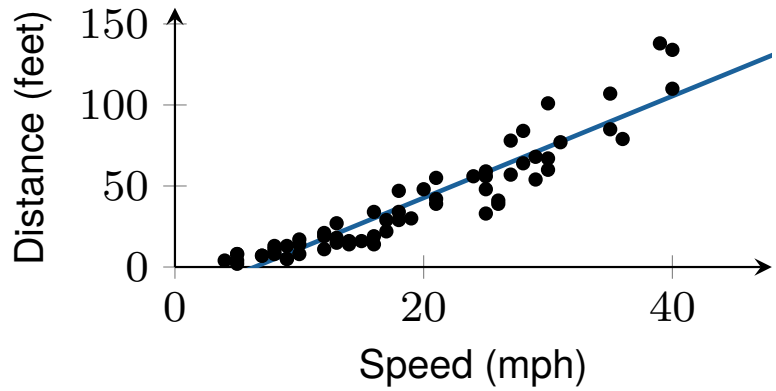
- x = Speed
- y = Distance

$$y = \beta_0 + \beta_1 x + \varepsilon, \;\; \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} 1 & 4 \\ 1 & 5 \\ 1 & 5 \\ 1 & 5 \\ 1 & 5 \\ 1 & 7 \\ 1 & 7 \\ 1 & 8 \\ \vdots & \vdots \\ 1 & 39 \\ 1 & 39 \\ 1 & 40 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} 4 \\ 2 \\ 4 \\ 8 \\ 8 \\ 7 \\ 7 \\ 8 \\ \vdots \\ 138 \\ 110 \\ 134 \end{bmatrix}$$

The normal equations $\Rightarrow \widehat{\beta} = \begin{bmatrix} -20.1 \\ 3.1 \end{bmatrix}$



Use the model for predictions!

Transforming the inputs

"If the speed v is an input variable, why can't the kinetic energy ($\propto v^2$) be an input variable?"

We can make arbitrary nonlinear transformations to the input variables!

The model is still a linear regression model, since

$$y = \beta_0 + \beta_1 x + \beta_2 v^2 + \beta_3 \cos(v) + \beta_4 \arctan(v) + \varepsilon$$

is equivalent to

$$\begin{aligned} y =& \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon, \\ \text{with} \quad X_1 &= v \\ X_2 &= v^2 \\ X_3 &= \cos(v) \\ X_4 &= \arctan(v) \end{aligned}$$

v = original input variable, x_i transformed input variables (features).

Example

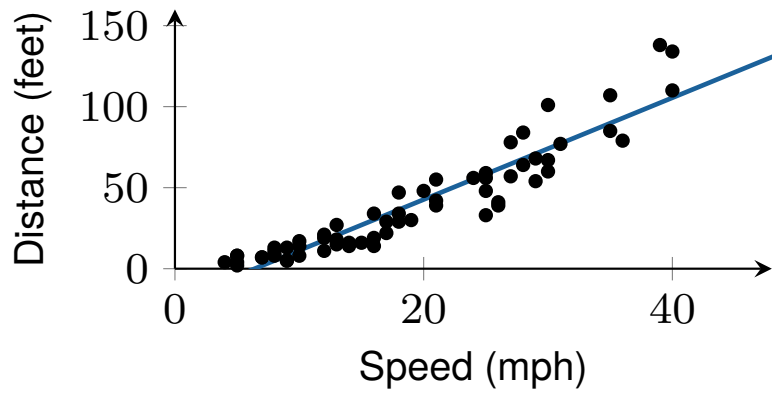
- x = Speed
- y = Distance

$$y = \beta_0 + \beta_1 x + \varepsilon, \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} 1 & 4 \\ 1 & 5 \\ 1 & 5 \\ 1 & 5 \\ 1 & 5 \\ 1 & 7 \\ 1 & 7 \\ 1 & 8 \\ \vdots & \vdots \\ 1 & 39 \\ 1 & 39 \\ 1 & 40 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} 4 \\ 2 \\ 4 \\ 8 \\ 8 \\ 7 \\ 7 \\ 8 \\ \vdots \\ 138 \\ 110 \\ 134 \end{bmatrix}$$

The normal equations $\Rightarrow \widehat{\beta} = \begin{bmatrix} -20.1 \\ 3.1 \end{bmatrix}$



Example

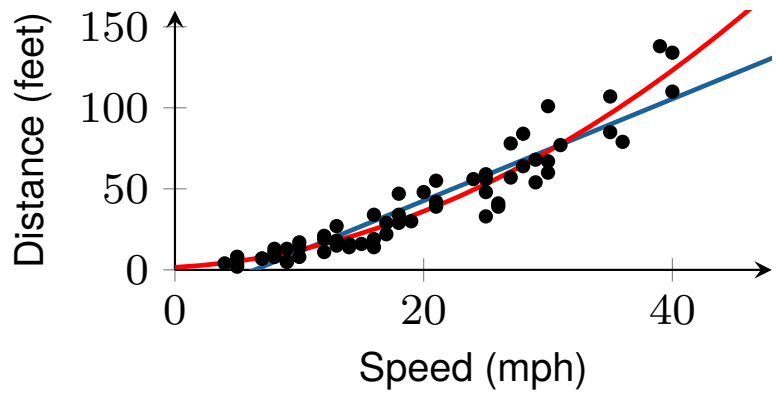
- x = Speed
- y = Distance

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon, \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} 1 & 4 & 16 \\ 1 & 5 & 25 \\ 1 & 5 & 25 \\ 1 & 5 & 25 \\ 1 & 5 & 25 \\ 1 & 7 & 49 \\ 1 & 7 & 49 \\ 1 & 8 & 64 \\ \vdots & \vdots & \vdots \\ 1 & 39 & 1521 \\ 1 & 39 & 1521 \\ 1 & 40 & 1600 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} 4 \\ 2 \\ 4 \\ 8 \\ 8 \\ 7 \\ 7 \\ 8 \\ \vdots \\ 138 \\ 110 \\ 134 \end{bmatrix}$$

The normal equations $\Rightarrow \widehat{\beta} = \begin{bmatrix} 1.58 \\ 0.42 \\ 0.066 \end{bmatrix}$



Transforming the inputs

If the original input variable is v, one can for instance use ...

- a polynomial in v

$$y = \beta_0 + \beta_1 \underset{\| \; x_1}{v} + \beta_2 \underset{\| \; x_2}{v^2} + \beta_3 \underset{\| \; x_3}{v^3} + \cdots + \beta_p \underset{\| \; x_p}{v^p} + \varepsilon$$

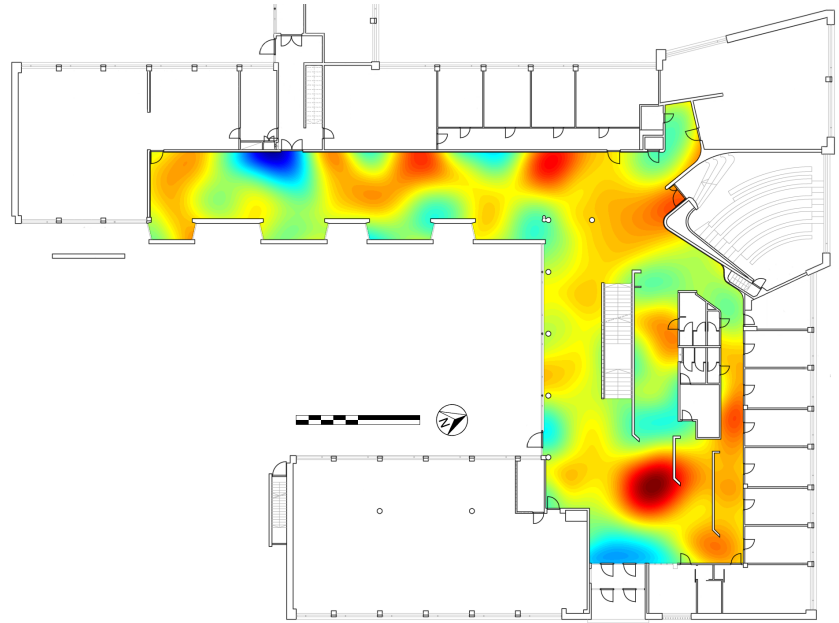
- radial basis function kernels (see lecture notes)
- ...

Ex) Ambient magnetic field map

The Earth's magnetic field sets a background for the ambient magnetic field. Deviations make the field vary from point to point.

Aim: Build a map (i.e., a model) of the magnetic environment based on magnetometer measurements.

Solution: Customized Gaussian process that obeys Maxwell's equations.



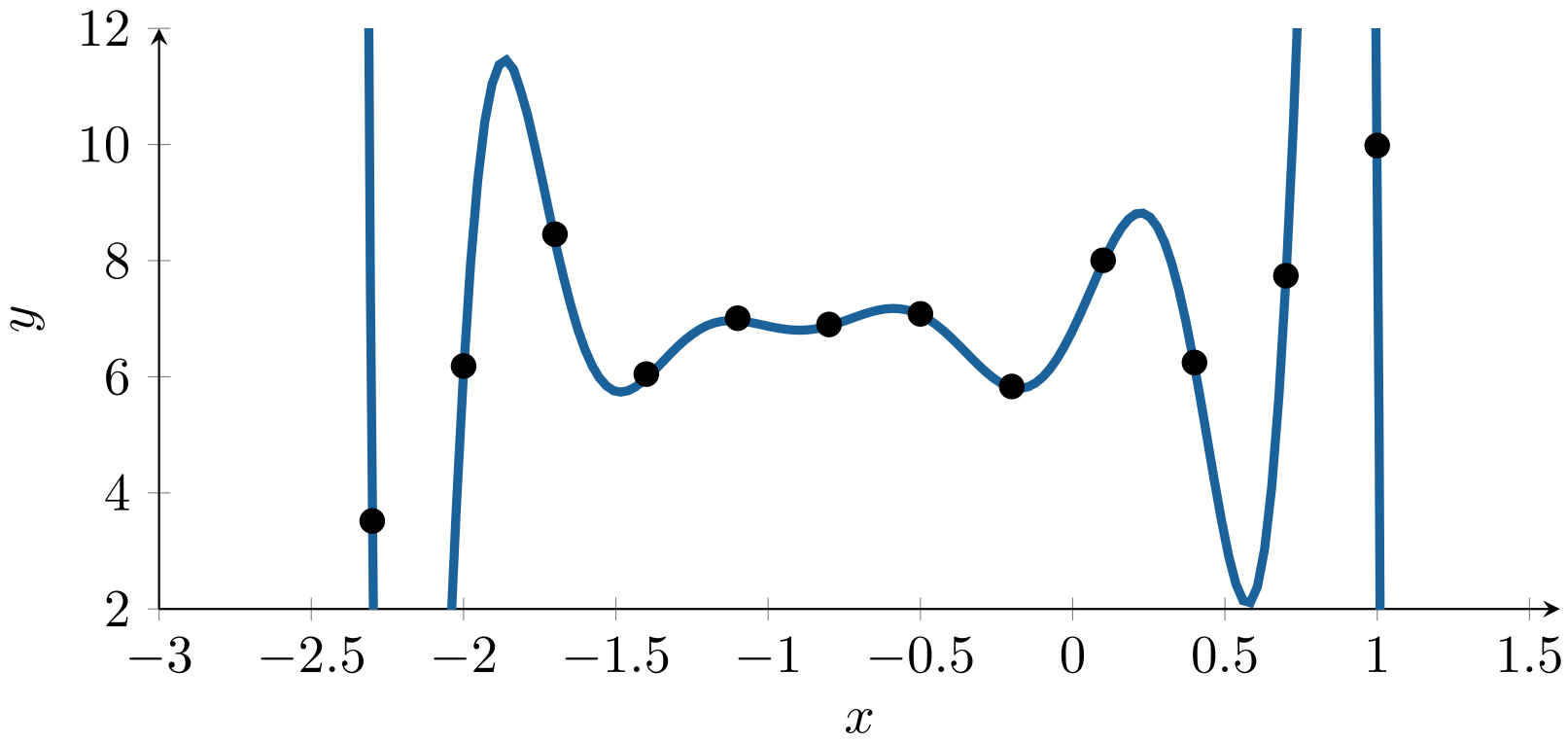
www.youtube.com/watch?v=enlMiUqPVJo

Arno Solin, Manon Kok, Niklas Wahlström, Thomas Schön and Simo Särkkä. **Modeling and interpolation of the ambient magnetic field by Gaussian processes**. *IEEE Transactions on Robotics*, 34(4):1112–1127, 2018.

Carl Jidling, Niklas Wahlström, Adrian Wills and Thomas Schön. **Linearly constrained Gaussian processes**. *Advances in Neural Information Processing Systems (NIPS)*, Long Beach, CA, USA, December, 2017.

A too flexible model?

With a $p = n - 1$ degree polynomial, we can fit n data points perfectly.



Is this desired? **Overfit!**

Regularization

"Keep β small unless the data really convinces us otherwise"

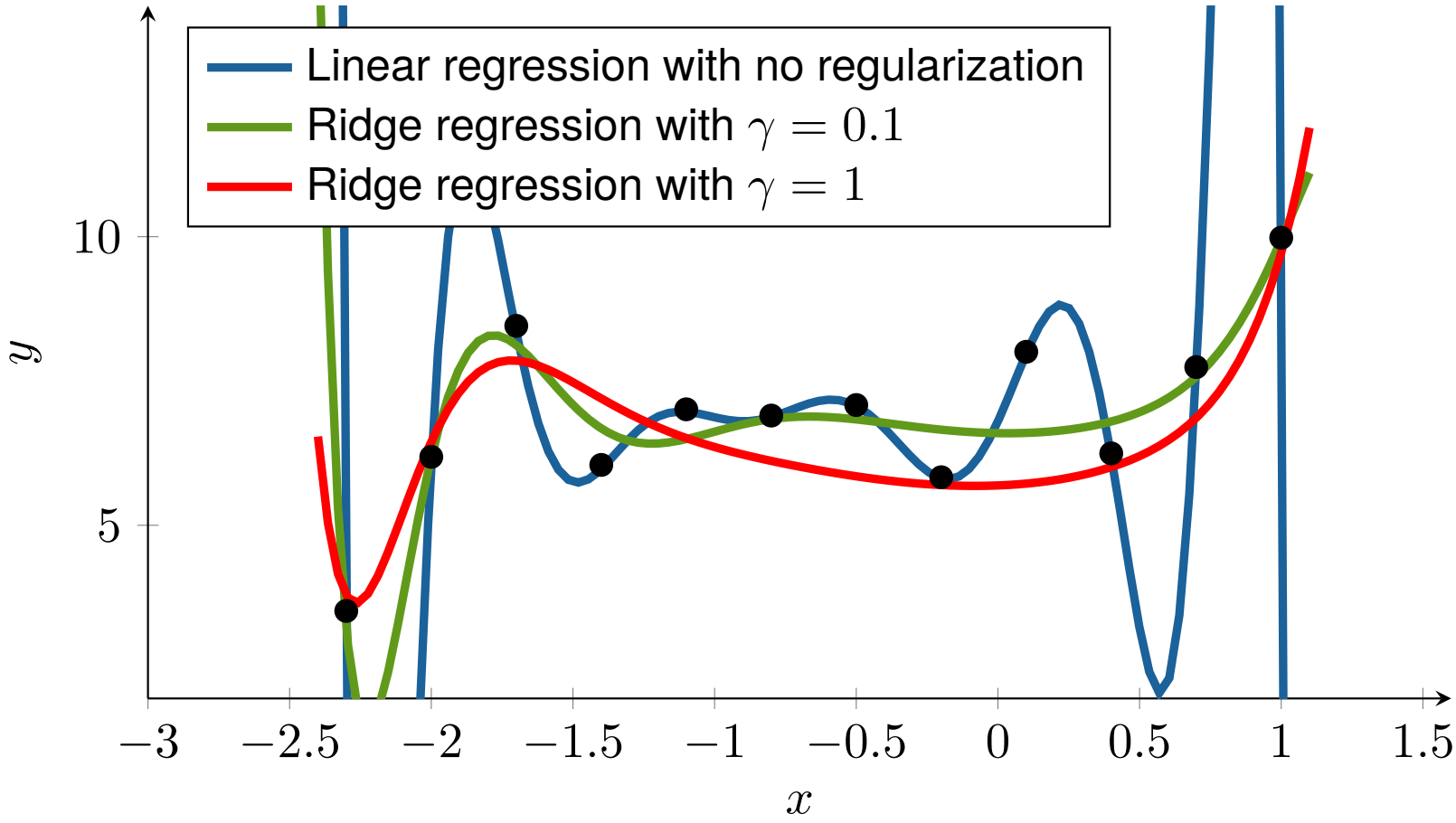
Least squares with Ridge regression

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \gamma\|\boldsymbol{\beta}\|_2^2$$

$$\Rightarrow (\mathbf{X}^\mathsf{T}\mathbf{X} + \gamma\mathbf{I}_{p+1})\widehat{\boldsymbol{\beta}} = \mathbf{X}^\mathsf{T}\mathbf{y}$$

γ regularization parameter

A too flexible model?



Regularization can help us to avoid overfitting!

Regularization

Ridge regression

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \gamma\|\boldsymbol{\beta}\|_2^2$$

(has a closed-form solution for $\widehat{\boldsymbol{\beta}}$)

LASSO

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \gamma\|\boldsymbol{\beta}\|_1$$

(lacks a closed-form solution for $\widehat{\boldsymbol{\beta}}$)

Regularization can be used in many methods, not only linear regression!

Dummy variables for qualitative inputs

For a qualitative input with 2 different classes/levels/labels A and B:
Create a dummy variable

$$x = \begin{cases} 0 & \text{if A} \\ 1 & \text{if B} \end{cases}$$

$$\Rightarrow y = \beta_0 + \beta_1 x + \varepsilon = \begin{cases} \beta_0 + \varepsilon & \text{if A} \\ \beta_0 + \beta_1 + \varepsilon & \text{if B} \end{cases}$$

Dummy variables for qualitative inputs

For a qualitative input with $k = 4$ different classes/levels/labels A, B, C, D:
Create $k - 1 = 3$ dummy variables

$$x_1 = \begin{cases} 1 & \text{if B} \\ 0 & \text{if not B} \end{cases}, \quad x_2 = \begin{cases} 1 & \text{if C} \\ 0 & \text{if not C} \end{cases}, \quad x_3 = \begin{cases} 1 & \text{if D} \\ 0 & \text{if not D} \end{cases}$$

$$\Rightarrow y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon = \begin{cases} \beta_0 + \varepsilon & \text{if A} \\ \beta_0 + \beta_1 + \varepsilon & \text{if B} \\ \beta_0 + \beta_2 + \varepsilon & \text{if C} \\ \beta_0 + \beta_3 + \varepsilon & \text{if D} \end{cases}$$

UPPSALA UNIVERSITET

A few concepts to summarize lecture 2

Regression is about learning a model that describes the relationship between an input variable $\mathbf{x}$ (quantitative or qualitative) and a quantitative output variable y.

Linear regression corresponds to regression with a linear model.

Maximum likelihood with Gaussian iid assumption on ε
$\Rightarrow$ **least squares** and **normal equations**

Nonlinear transformations can be applied to the input variables

Overfit is when the model adapts (too much) to noise in the data

Regularization can help against overfitting

Qualitative variables are handled by dummy variables