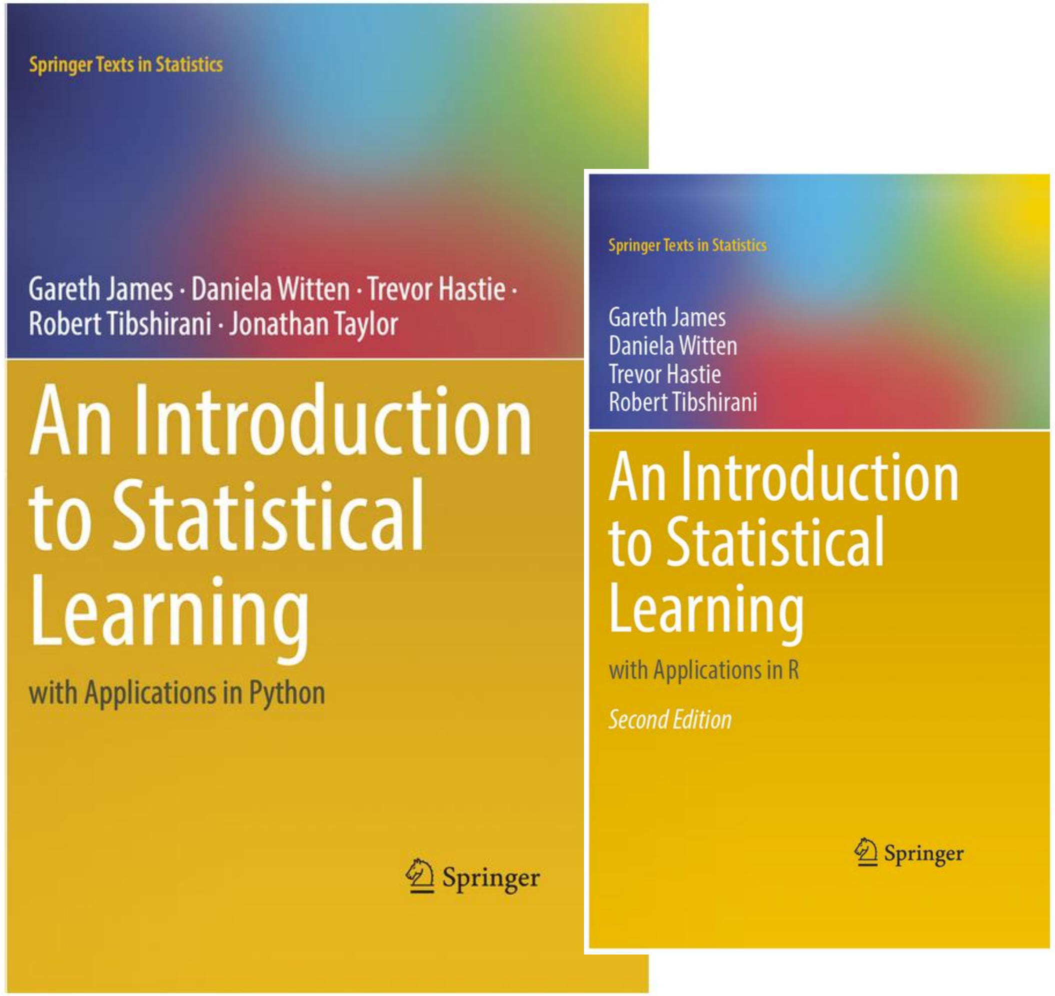
Statistical learning and linear regression

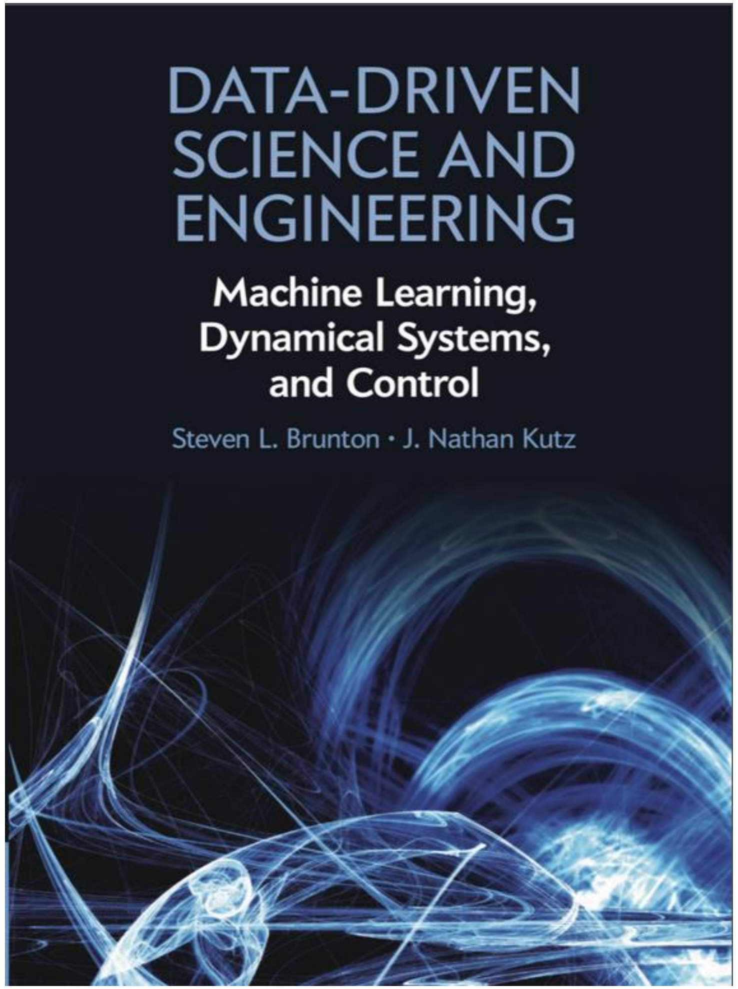
Lecture 5 of “Mathematics and AI”

Outline

1. Supervised learning
2. Linear regression
3. Linear regression on multiple variables
4. Strengths and limitations

Reading on statistical learning





Supervised learning

Supervised learning

Hello Machine …

Let me show you some queries …

Let me tell you the correct answers to those queries …

Find the pattern!

Here are some queries that
you haven't seen before.

Let me check how well you can answer those based on the pattern that you learned.

(What is the capitol of France?, Paris)

( , "Cat")

$(x = 0.2, y = 1.3)$

Image source: Photo by Dr. Schwarze of Dr. Schwarze's cat Emmy

Supervised learning

Hello Machine …

Let me show you some queries …

Let me tell you the correct answers to those queries …

Find the pattern!

Here are some queries that
you haven't seen before.

Let me check how well you can answer those based on the pattern that you learned.

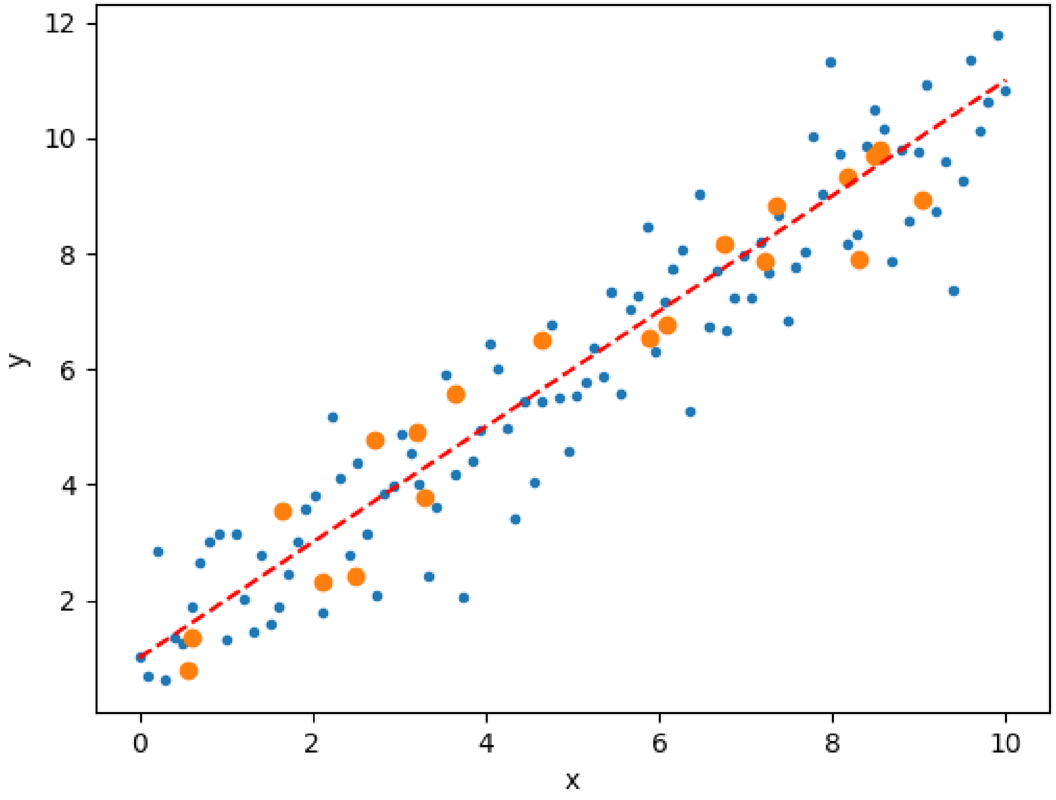
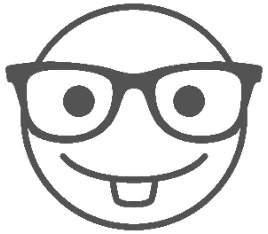


Image source: Photo by Dr. Schwarze of Dr. Schwarze's cat Emmy

Supervised learning

	Statistics	Machine learning
Hello Machine …		
Let me show you some queries … Let me tell you the correct answers to those queries …	Sample	Training set
Find the pattern!	Fit the model Quality of fit (within sample)	Train a model Training accuracy
Here are some queries that you haven’t seen before.	Out-of-sample prediction	Test set
Let me check how well you can answer those based on the pattern that you learned.	Out-of-sample quality of fit	Test accuracy

Linear regression

Linear regression: sample / training set

Sample of size n is a set of n value pairs:

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

We can the data in two vectors:

$$(x_1, x_2, \dots, x_n), (y_1, y_2, \dots, y_n)$$

Linear regression: sample / training set

Sample of size n is a set of n value pairs:

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

We can the data in two vectors:

$$(x_1, x_2, \dots, x_n), \quad \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Linear regression: the model

Model parameters

Independent variable
(query, input, feature)

$$Y \approx \beta_o + \beta_1 X$$

Dependent variable
(response, output)

Intercept

Slope

Linear regression: fitting the model

Estimated
model parameters

$$Y \approx \hat{\beta}_o + \hat{\beta}_1 X$$

Estimated
intercept

Estimated
slope

Model parameters are unknown. Need to be estimated from data.

Linear regression: fitting the model

In general, model fitting can involve:

- complex training algorithms
- many iterations of (re-)estimating model
  parameters and
- assessing the quality of fit to the training data.

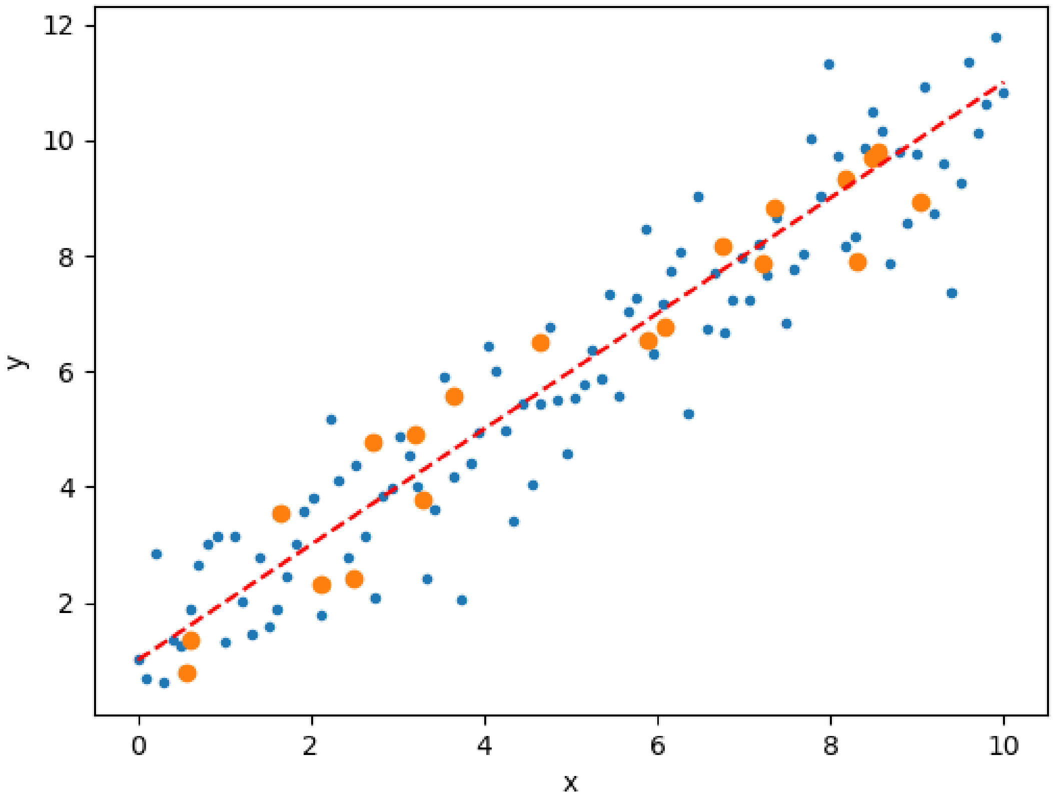
Not for linear regression.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Linear regression: Quality of fit

How well does our line fit the data?



Residual sum of squares (RSS)

$$RSS = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

Linear regression: Quality of fit

Total sum of squares (TSS)

$$TSS = \sum_{i=1}^{n} (\bar{y}_i - y_i)^2$$

RSS of constant model

Residual sum of squares (RSS)

$$RSS = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

account for sample size and “model complexity”

Residual standard error (RSE)

$$RSE = \sqrt{\frac{RSS}{n-2}}$$

account for sample size

Mean squared error (MSE)

$$MSE = \frac{RSS}{n}$$

Fraction of variance explained compared to constant model

Variance explained (R^2)

$$R^2 = \frac{TSS - RSS}{TSS}$$

Linear regression: quality of fit

What does it mean when the quality of fit is low?

- Bad parameter estimation
    - not an issue for linear regression
- Relationship too weak or data set too small
    - check via significance test: t statistic, p value
- Bad model
    - e.g. “very” non-linear relationship between x and y

Multivariate linear regression

Multivariate linear regression: the model

Independent variables
(query, input, feature)

$$Y \approx \beta_o + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Dependent variable
(response, output)

Intercept

Slopes

Multivariate linear regression: fitting the model

$$Y \approx \hat{\beta}_o + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p$$

Model parameters are unknown. Need to be estimated from data.

Multivariate linear regression: fitting the model

For centered data:

$$\begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_p \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Multivariate linear regression: fitting the model

For centered data:

$$X\hat{\beta} = Y$$

Matrix of inputs

Vector of parameters

Vector of outputs

Strengths and limitations of linear regression

Strengths

1. Simple model
2. Simple “training procedure”
3. “Convergence” guaranteed
4. Optimality guaranteed (see Gauss-Markov theorem)
5. Well-established quality of fit measures
6. Good starting point for regression problems

Limitations

- Non-linearity of the response-predictor relationships
  - ➢ Residual plots can help identify non-linearity.
- Correlation of error terms
  - ➢ Can lead to inefficient estimates of the coefficients.
- Heteroscedasticity: Non-constant variance of error terms
  - ➢ violates the assumptions of the linear regression model.
- Outliers
  - ➢ Points that have a large influence on the fit of the model.
- Collinearity:
  - ➢ When predictor variables are highly correlated, it's difficult to separate out the individual effects of each predictor.