DATA11002 Introduction to Machine Learning

Kai Puolamäki

2 December 2020

Announcements

- Machine Learning Guest Lectures on **4 December** (poster)
  - Andreas Henelius (OP Financial Group): *Data Science in Finance - Machine learning for overdue invoice prediction*
  - Antti Ukkonen (Speechly): *From voice to meaning: Machine learning for spoken language understanding*
  - Discussion
- Term project and E3 DL on **6 December**
- Max. 2 min. video pitch on **9 December** (replaces 9 Dec lecture, which is canceled)
- Selected term project presentations on **11 December**
- Term project final report on **20 December**

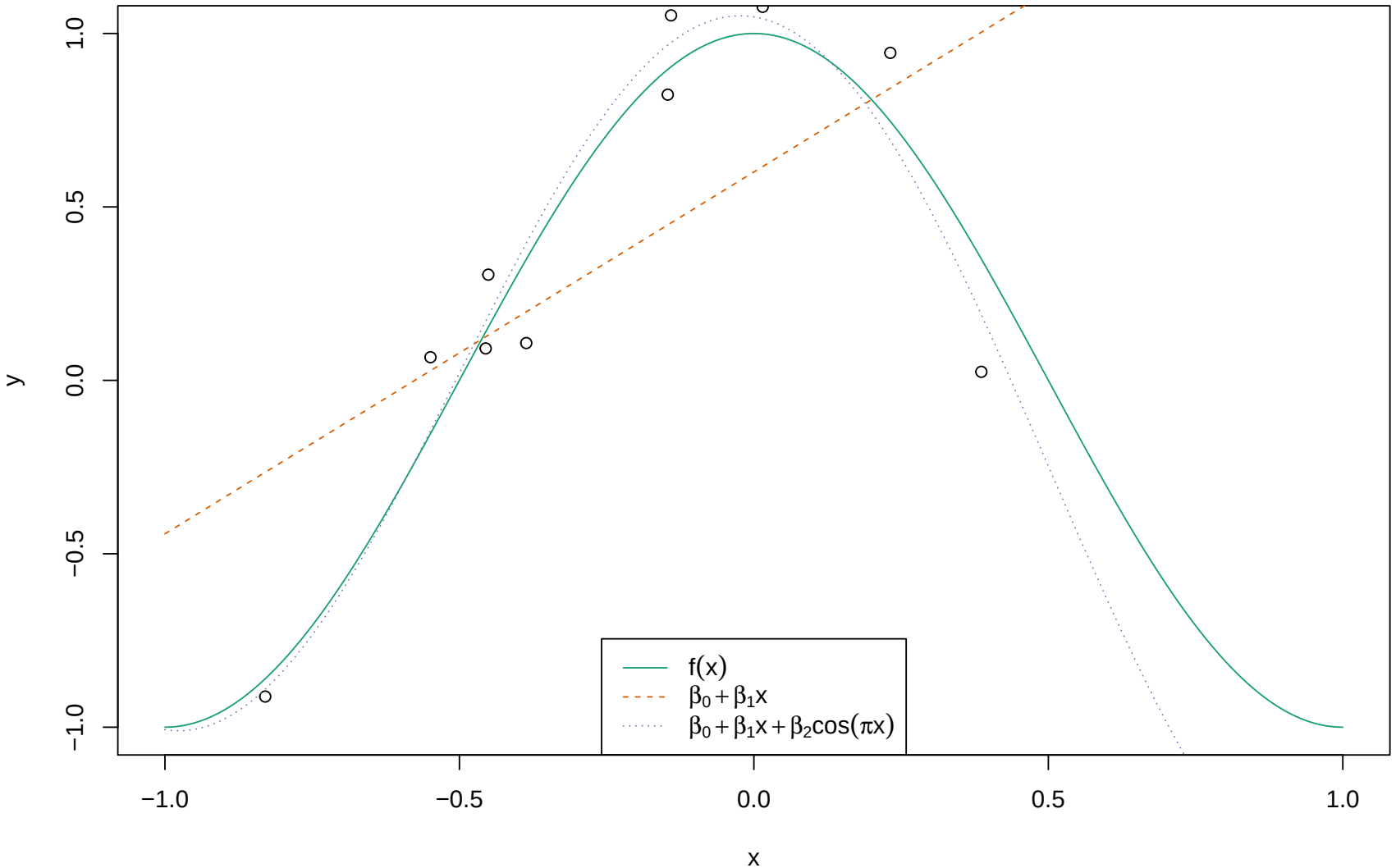
Recap and practical considerations

Random thoughts

- 42
- Typical data analysis process: what is interesting and/or difficult
- When the reality hits and the assumptions break
  - typical work-flow
  - the real meaning of our assumptions
    - iid vs time series data
    - concept drift
    - common confusions
  - other assumptions that go typically wrong
  - how to tackle these in practice

Statistics is often misleading

- You have to be careful in interpreting any significance results
- Is there an increasing (decreasing) trend?



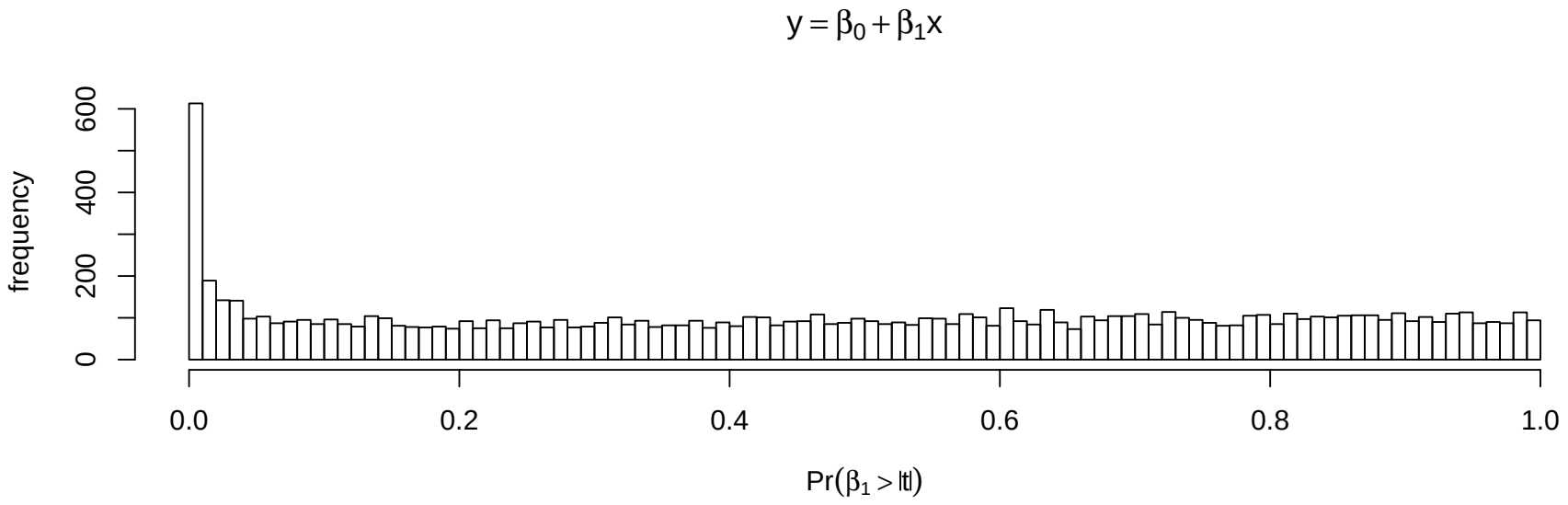
Statistics is often misleading

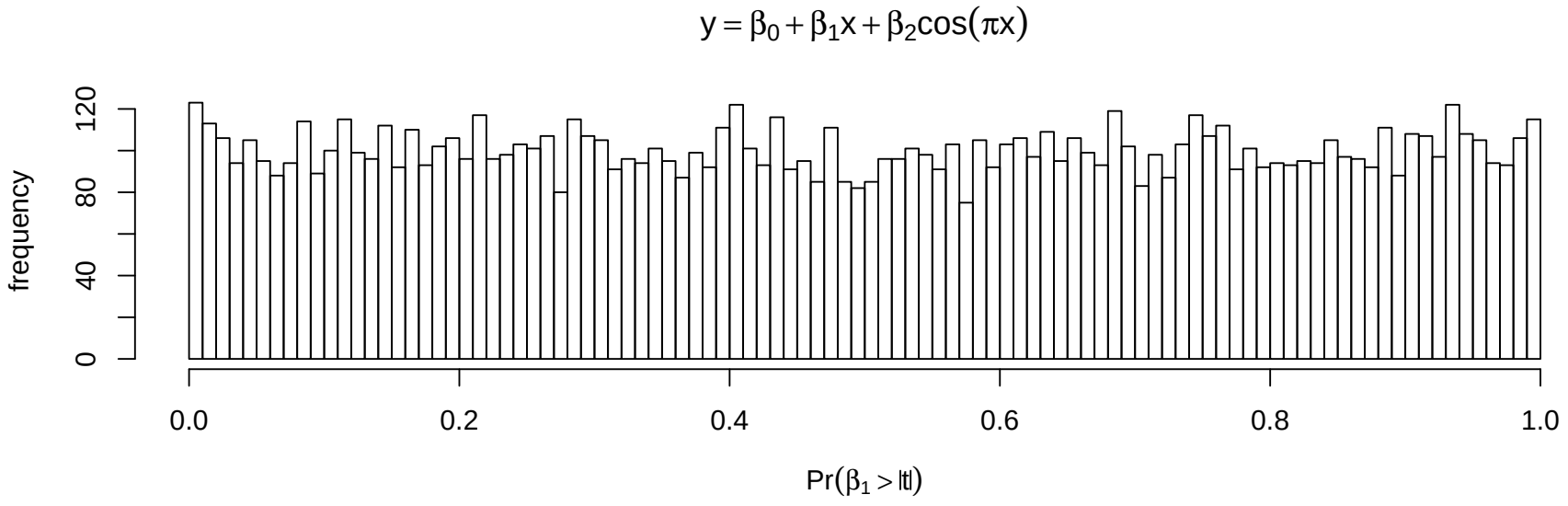
```
##
## Call:
## lm(formula = y ~ x, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.97901 -0.07587  0.07028  0.32519  0.59826
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.6007     0.1956   3.070   0.0153 *
## x             1.0433     0.4619   2.259   0.0538 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## Residual standard error: 0.5171 on 8 degrees of freedom
## Multiple R-squared:  0.3894, Adjusted R-squared:  0.3131
## F-statistic: 5.102 on 1 and 8 DF,  p-value: 0.05383
```

Statistics is often misleading

```
##
## Call:
## lm(formula = y ~ x + cos(pi * x), data = data)
##
## Residuals:
##       Min         1Q     Median         3Q        Max
## -0.287964 -0.126945  0.006543  0.108892  0.252524
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.1135     0.1260  -0.901 0.397561
## x            -0.2678     0.2560  -1.046 0.330228
## cos(pi * x)   1.1614     0.1659   7.002 0.000211 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## Residual standard error: 0.1954 on 7 degrees of freedom
## Multiple R-squared:  0.9237, Adjusted R-squared:  0.9019
## F-statistic: 42.38 on 2 and 7 DF,  p-value: 0.0001226
```

Statistics is often misleading





Statistics is often misleading

- What just happened?
- We just basically ruled out the null hypothesis that the data comes from a model family specified by $y = \beta_0 + \epsilon$.
  - this does not allow us to conclude that there is an increasing trend!
- When the model is correct then the "trend" disappears, i.e., we cannot rule out the null hypothesis that $\beta_1 = 0$ in a model family given by $y = \beta_0 + \beta_1 x + \beta_2 \cos(\pi x) + \epsilon$.
- This is why it is important to check for normality, homoscedasticity etc.
  - if the data is from "outside the model" (there are no parameters that would really describe the model) then there is only so much that we can conclude about the parameter values

Statistics can be misleading

- Other issues:
  - with large enough data (almost) everything is significant
  - statistically significant vs. practically relevant
  - Woolston (2015) Psychology journal bans P values. Nature.

Dealing with partially broken assumptions

- Often it is useful to assume something even if we know that it is not 100% accurate (in natural sciences you do this all the time!)
- It is important to look at what you did and if your assumptions are true (not restricted to looking at your residual distributions etc.!)
  - the most major assumptions are often related to the models and variables chosen (which are implicit and not often thought as assumptions); e.g., if you use linear models then you make a very strong assumptions that your data is linear (normality etc. assumptions may be quite minor compared to this!)

Example: autocorrelation and time series data

- Consider time series regression setup:

$$\ldots, z_{-2}, z_{-1}, z_0, z_1, z_2, \ldots$$

  - denote subsequence by $z_{ab} = (z_a, \ldots, z_b)$, where $a \leq b$.
- Relevant properties:
  - i.i.d.: typically broken, $P(z_a, \ldots, z_b) \neq P(z_a) \times \ldots \times P(z_b)$ (autocorrelation).
  - (strong) stationarity; $P(z_{ab}) = P(z_{a+\tau, b+\tau})$ is true for all time intervals $[a, b]$ and time shifts τ.
    - homoscedasticity: $\mathrm{var}(z_a) = \mathrm{var}(z_b)$ for all a, b.
    - weak stationarity: mean is constant and autocorrelation depends only on lag.
  - ergodicity: $\lim_{n\to\infty} |E[f(z_{ab})]E[g(z_{a+n,b+n})]| = |E[f(z_{ab})]|\,|E[g(z_{a+n,b+n})]|$.
    - intuition: after sufficiently long time time time series "forgets" earlier state and can end up in any allowed state with non-zero probability.

Example: autocorrelation and time series data

- If time series is stationary and ergodic but not i.i.d. (=autocorrelation exists):
  - estimator of mean of z is still unbiased
  - variance of the estimator of mean may be larger ("effective sample size")
  - intuition: learned ML model may be correct on average, but it may overfit
- i.i.d. may be too strong requirement for regression (classification)
  - $y_t = f(x_t) + \epsilon_t$
  - it is enough that the residuals ϵ_t are i.i.d.
- Use of validation set methods:
  - it is usually enough if validation set is independent of the training set
  - E.g., train your model data using earlier observations and validate using later observations (and hope that these sets are approximately independent)

Example: autocorrelation and time series data

- Often doing machine learning on non-i.i.d. data works and gives the correct models on expectation
- Typically the confidence estimates may however be off, leading the overfitting (if autocorrelation is not taken into account)
- Danger: overcompensating time series effect may lead to undesirable results
  - E.g., detrending the data without thinking may remove some autocorrelation but it may also remove the effect of interest (the interesting effect may be autocorrelated as well!)
- In all machine learning: our models and assumptions are almost never fully satisfied
  - it is important to understand what is broken and what are the consequences.