COMPSCI762: Foudnations of Machine Learning

Regression

Jörg Simon Wicker
The University of Auckland

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

SCIENCE
SCHOOL OF COMPUTER SCIENCE
MACHINE LEARNING

Today we will cover...

Regression

Linear Regression

Least Squares

Different Notation

Summary

Partially based on Slides from University of British Columbia

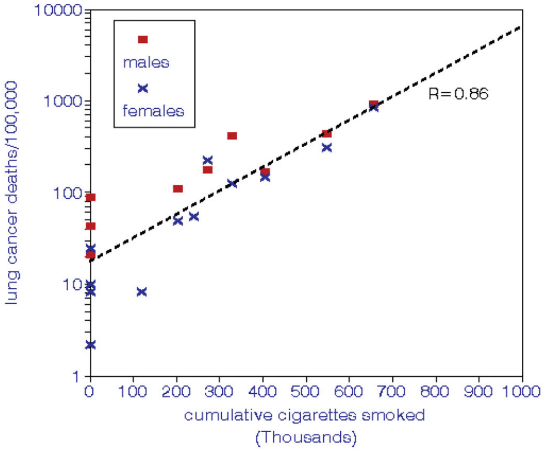
Regression

Supervised Learning Round 2: Regression

- We are going to revisit supervised learning
- Previously, we considered classification
  - We assumed y_i was discrete: $y_i = spam$ or $y_i = not\ spam$
- Now we are going to consider regression
  - We allow y_i to be numerical, for example $y_3 = 10.34cm$

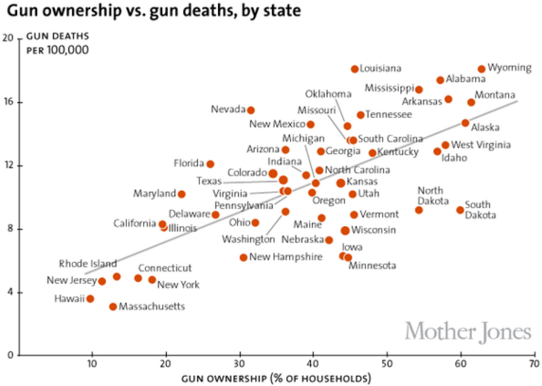
Example: Dependent vs. Explanatory Variables

- We want to discover relationship between numerical variables
  - Does number of lung cancer deaths change with number of cigarettes?
  - Does number of skin cancer deaths change with latitude?
  - Do people in big cities walk faster?
  - Is the universe expanding or shrinking or staying the same size?
  - Does number of gun deaths change with gun ownership?
  - Does number violent crimes change with violent video games?



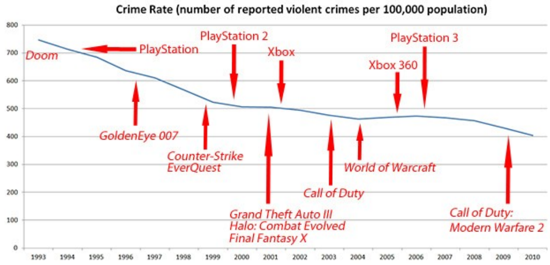
Example: Dependent vs. Explanatory Variables

- We want to discover relationship between numerical variables
  - Does number of lung cancer deaths change with number of cigarettes?
  - Does number of skin cancer deaths change with latitude?
  - Do people in big cities walk faster?
  - Is the universe expanding or shrinking or staying the same size?
  - Does number of gun deaths change with gun ownership?
  - Does number violent crimes change with violent video games?



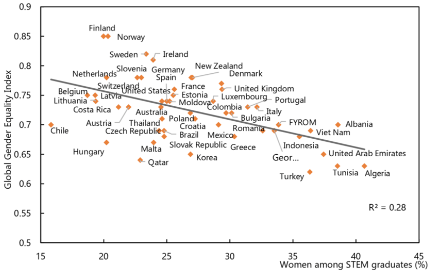
Example: Dependent vs. Explanatory Variables

- We want to discover relationship between numerical variables
  - Does number of lung cancer deaths change with number of cigarettes?
  - Does number of skin cancer deaths change with latitude?
  - Do people in big cities walk faster?
  - Is the universe expanding or shrinking or staying the same size?
  - Does number of gun deaths change with gun ownership?
  - Does number violent crimes change with violent video games?



Example: Dependent vs. Explanatory Variables

- We want to discover relationship between numerical variables
  - Does higher gender equality index lead to more women STEM grads?
- Not that we're doing supervised learning
  - Trying to predict value of 1 variable (the y_i values) – instead of measuring correlation between 2
- Supervised learning does not give causality
  - OK: Higher gender equality index is correlated with lower graduation rate
  - OK: Higher gender equality index helps predict lower graduation rate
  - BAD: Higher gender equality index leads to lower graduation rate



Handling Numerical Labels

- One way to handle numerical y_i: discretize
  - E.g., for 'age' could we use $age \leq 20$, $20 < age \leq 30$, $age > 30$
  - Now we can apply methods for classification to do regression
  - But coarse discretization loses resolution
  - And fine discretization requires lots of data
- There exist regression versions of classification methods:
  - Regression trees, probabilistic models, non-parametric models
  - Today: one of oldest, but still most popular/important methods
    - Linear regression based on squared error
    - Interpretable and the building block for more-complex methods

Linear Regression

Linear Regression in 1 Dimension

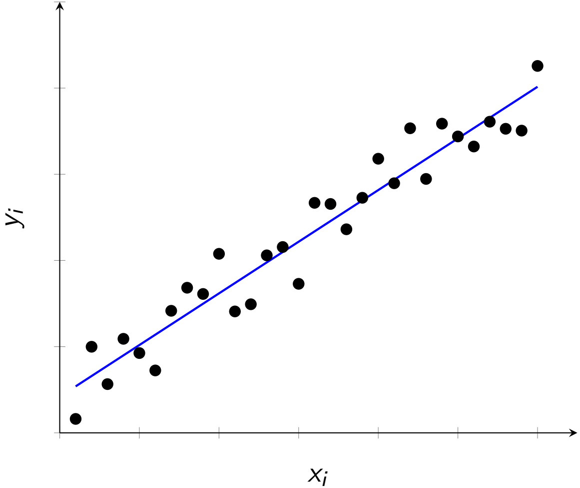
SCIENCE
SCHOOL OF COMPUTER SCIENCE
MACHINE LEARNING

- Assume we only have 1 feature ($d = 1$)
  - E.g., x_i is number of cigarettes and y_i is number of lung cancer deaths
- Linear regression makes predictions $\hat{y}_i$ using a linear function of x_i

$$\hat{y}_i = wx_i$$

- The parameter w is the weight or regression coefficient of x_i
  - We are temporarily ignoring the y-intercept
- As x_i changes, slope w affects the rate that $\hat{y}_i$ increases/decreases
  - Positive w: $\hat{y}_i$ increase as x_i increases
  - Negative w: $\hat{y}_i$ decreases as x_i increases

Linear Regression in 1 Dimension



line $\hat{y}_i = wx_i$ for a particular slope w

Least Squares

Least Squares Objective

- Our linear model is given by
$$\hat{y}_i = wx_i$$
- So we make predictions for a new example by using
$$\hat{y}_i = w\tilde{x}_i$$
- But we can't use the same error as before
  - It is unlikely to find a line where $\hat{y}_i = y_i$ exactly for many points
    - Due to noise, relationship not being quite linear or just floating-point issues
  - Best model may have $|\hat{y}_i - y_i|$ is small but not exactly 0

Least Squares Objective

- Instead of *exact* y_i, we evaluate size of the error in prediction
- Classic way is setting slope w to minimize sum of squared errors

Prediction $\hat{y}_i$ True value of y_i

$$f(w) = \sum_{i=1}^{n} (wx_i - y_i)^2$$

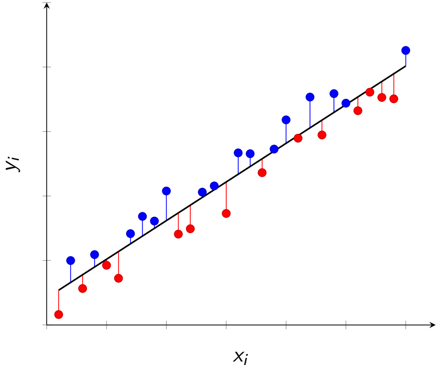
Sum over all training examples

Squared difference between prediction and true value for example x_i

- There are some justifications for this choice
  - A probabilistic interpretation is coming later in the course
- But usually, it is done because it is easy to minimize

Least Squares Objective

- Classic way to set slope w is minimizing sum of squared errors
$$f(w) = \sum_{i=1}^{n}(wx_i - y_i)^2$$



- "Error" is the sum of the squared values of these vertical distances between the line (wx_i) and the targets (y_i)
- If this error is small then our predictions are close to the target

Minimizing a Differential Function

- Simple approach to minimizing a differentiable function f
  1. Take the derivative of f
  2. Find points w where the derivative $f'(w)$ is equal to 0
  3. Choose the smallest one (and check that $f''(w)$ is positive).
- Note that this problem: $f(w) = \sum_{i=1}^{n}(wx_i - y_i)^2$
- Has the same set of minimizers as this problem: $f(w) = \frac{1}{2}\sum_{i=1}^{n}(wx_i - y_i)^2$
- And these also have the same minimizers: $f(w) = \frac{1}{n}\sum_{i=1}^{n}(wx_i - y_i)^2$, $f(w) = \frac{1}{2n}\sum_{i=1}^{n}(wx_i - y_i)^2 + 1000$
- We can multiply f by any positive constant and not change solution
  - Derivative will still be zero at the same locations
  - We will use this trick a lot!

Finding Least Squares Solution

- Finding w that minimizes the sum of squared errors

$$
\begin{aligned}
f(w) = \frac{1}{2}\sum_{i=1}^{n}(wx_i - y_i)^2 &= \frac{1}{2}\sum_{i=1}^{n}\left[w^2x_i^2 - 2wx_iy_i + y_i^2\right] \\
&= \frac{w^2}{2}\sum_{i=1}^{n}x_i^2 - w\sum_{i=1}^{n}x_iy_i + \frac{1}{2}\sum_{i=1}^{n}y_i^2 \\
&= \frac{w^2}{2}a - wb + c \\
\mathit{Take\ derivative}\ f'(w) &= wa - b + 0
\end{aligned}
$$

Setting $f'(w) = 0$ and solving gives

$$
w = \frac{b}{a} = \frac{\sum_{i=1}^{n} x_iy_i}{\sum_{i=1}^{n} x_i^2}
$$

Finding Least Squares Solution

- Finding w that minimizes sum of squared errors

$$w = \frac{b}{a} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

- Let's check that this is a minimizer by checking the second derivative

$$f'(w) = w \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i y_i$$

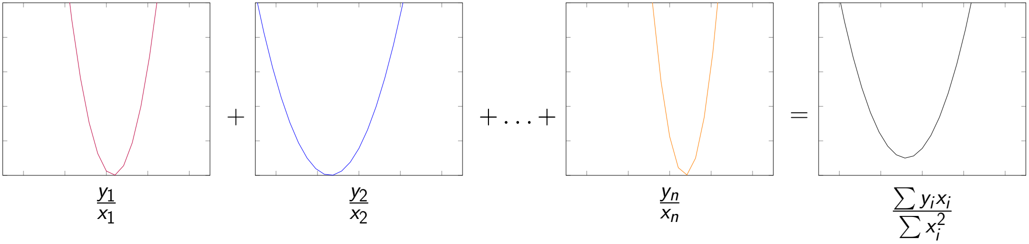
$$f''(w) = \sum_{i=1}^{n} x_i^2$$

  - Since $(anything)^2$ is non-negative and $(anything\ non - zero)^2 > 0$, if we have one non-zero feature then $f''(w) > 0$ and this is a minimizer

Least Squares Objective / Solution (Another View)

- Least squares minimizes a quadratic that is a sum of quadratics

$$f(w) = (wx_1 - y_1)^2 + (wx_2 - y_2)^2 + \ldots + (wx_n - y_n)^2$$



Motivation: Combining Explanatory Variables

- Smoking is not the only contributor to lung cancer
  - For example, there environmental factors like exposure to asbestos
- How can we model the combined effect of smoking and asbestos?
- A simple way is with a 2-dimensional linear function

$$\hat{y} = w_1 x_{i1} + w_2 x_{i2}$$

- We have a weight w_1 for feature 1 and w_2 for feature 2

$$\hat{y}_i = 10(\#cigarettes) + 25(\#asbestos)$$

Different Notation

Different Notations for Least Squares

- If we have d features, the d-dimensional linear model is

$$\hat{y}_i = w_1 x_{i1} + w_2 x_{i2} + \ldots + w_d x_{id}$$

  - In words, the output of our model is a weighted sum of the inputs
- We can re-write this in summation notation

$$\hat{y}_i = \sum_{j=1}^{d} w_j x_{ij}$$

- We can also re-write this in vector notation

$$\hat{y}_i = w^T x_i$$

Notation

- In my lectures, all vectors are assumed to be column-vectors

$$w = \begin{bmatrix} w_1 \\ w_2 \\ \cdots \\ w_d \end{bmatrix} \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix} \quad x_i = \begin{bmatrix} x_{i_1} \\ x_{i_2} \\ \cdots \\ x_{i_d} \end{bmatrix}$$

- So $w^T x_i$ is a scalar

$$w^T x_i = \begin{bmatrix} w_1 & w_2 & \cdots & w_d \end{bmatrix} \begin{bmatrix} x_{i1} \\ x_{i2} \\ \cdots \\ x_{id} \end{bmatrix} = w_1 x_{i1} + w_2 x_{i2} + \ldots + w_d x_{id} = \sum_{j=1}^{d} w_j x_{ij}$$

- So rows of X are actually transpose of column-vector x

$$X = \begin{bmatrix} x_1^T \\ x_2^T \\ \cdots \\ x_n^T \end{bmatrix}$$

Least Squares in d-Dimensions

- The linear least squares model in d-dimensions minimizes

$$f(w) = \frac{1}{2}\sum_{i=1}^{n}(wx_i - y_i)^2$$

  - w is now a vector
  - $w^T x_i$ (prediction) is inner product of w and x_i (linear combination of features)
  - $\sum_{i=1}^{n}(wx_i - y_i)^2$ (error) is still the sum of squared differences between true y_i and our prediction $w^T x_i$
- Dates back to 1801: Gauss used it to predict location of Ceres
- How do we find the best vector w in d dimensions?
  - Can we set the partial derivative of each variable to 0

Summary

Summary

- Regression considers the case of a numerical y_i
- Least squares is a classic method for fitting linear models
- With 1 feature, it has a simple closed-form solution
- Can be generalized to d features
  - What does the regression look like in 2 dimensions?
- There are many more regression models
  - Model trees, regression trees

Literature

- Machine Learning – Tom Mitchell
- Pattern Recognition and Machine Learning – Christopher Bishop
- Data Mining – Jiawei Han, Micheline Kamber, Jian Pei
- Data Mining – Ian Witten, Eibe Frank, Mark Hall, Christopher Pal

SCIENCE
SCHOOL OF COMPUTER SCIENCE
MACHINE LEARNING

Thank you for your attention!

https://ml.auckland.ac.nz