CPSC 340: Machine Learning and Data Mining

Linear Least Squares

Fall 2015

Admin

- Assignment 3 out today.
  - Longer than other assignments, but due on October 23rd.
- Midterm moved to October 30.
  - Covers Assignments 1-3.
  - Practice midterm coming.

User-Product Matrix

Column x^j gives all features x_{ij} for one product 'j'.

If $x_{ij} = 1$, means user 'i' bought item 'j'.

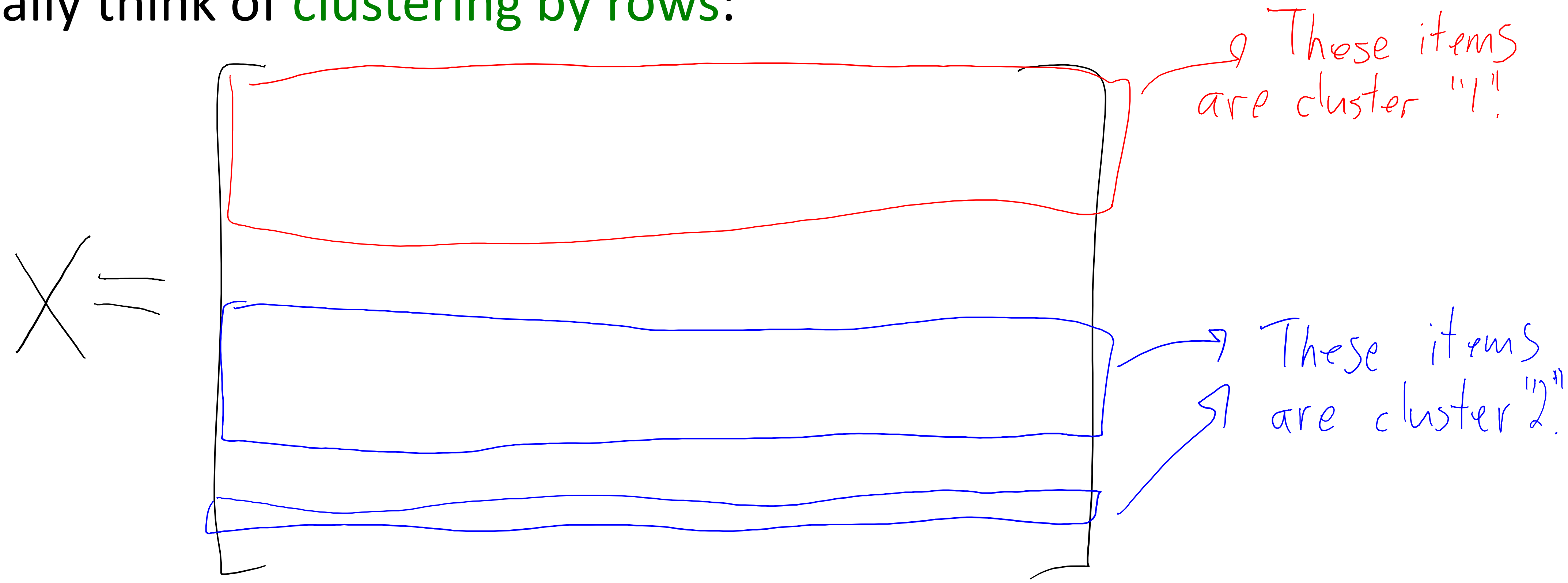
$$X = \begin{bmatrix} & & 1 \\ & & \\ & & 0 \end{bmatrix}$$

If $x_{ij} = 0$, means user 'i' has not bought item 'j'.

Row x_i gives all features x_{ij} for one user 'i'. By convention, x_i is a $d \times 1$ column-vector.

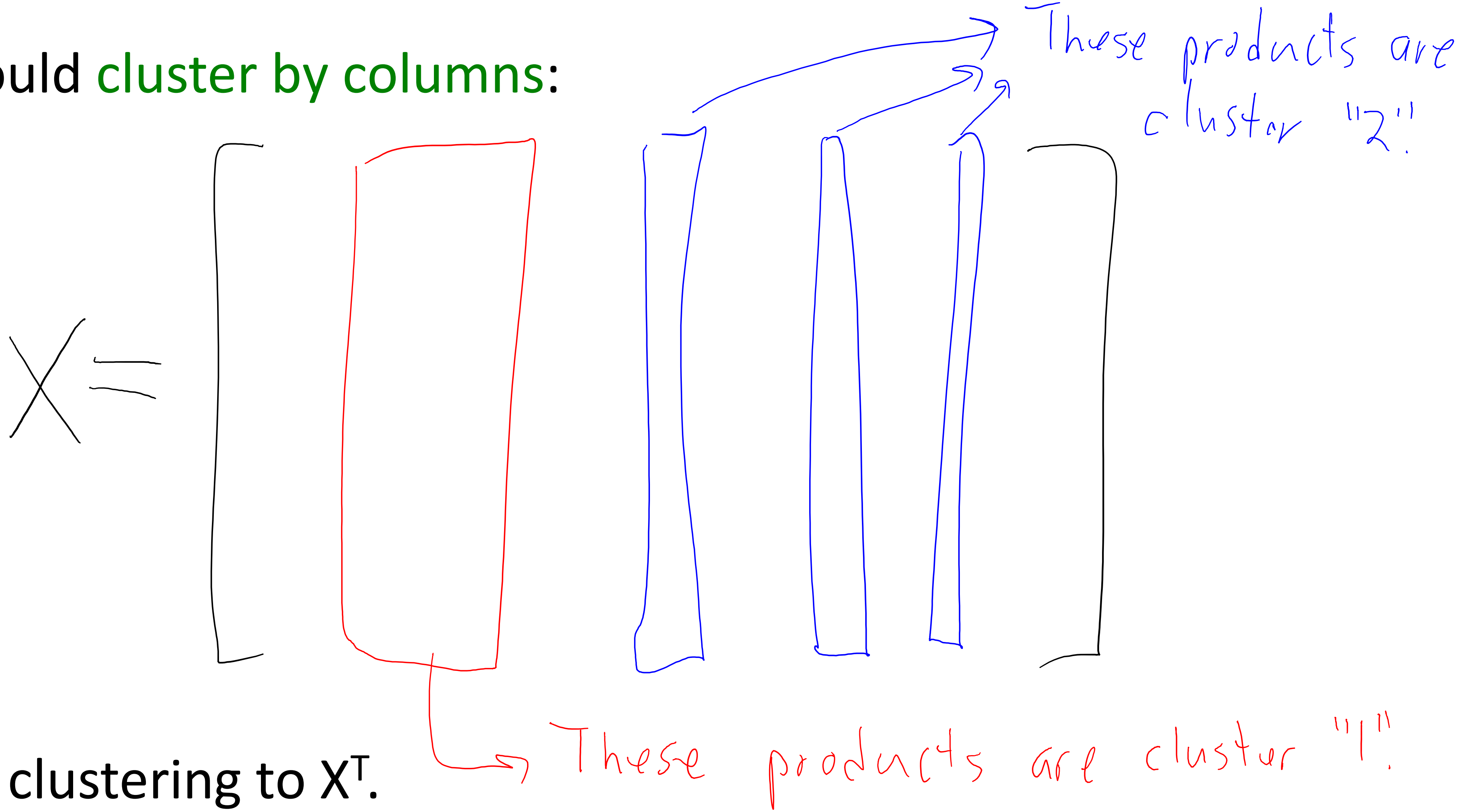
Clustering User-Product Matrix

- Normally think of clustering by rows:



Clustering User-Product Matrix

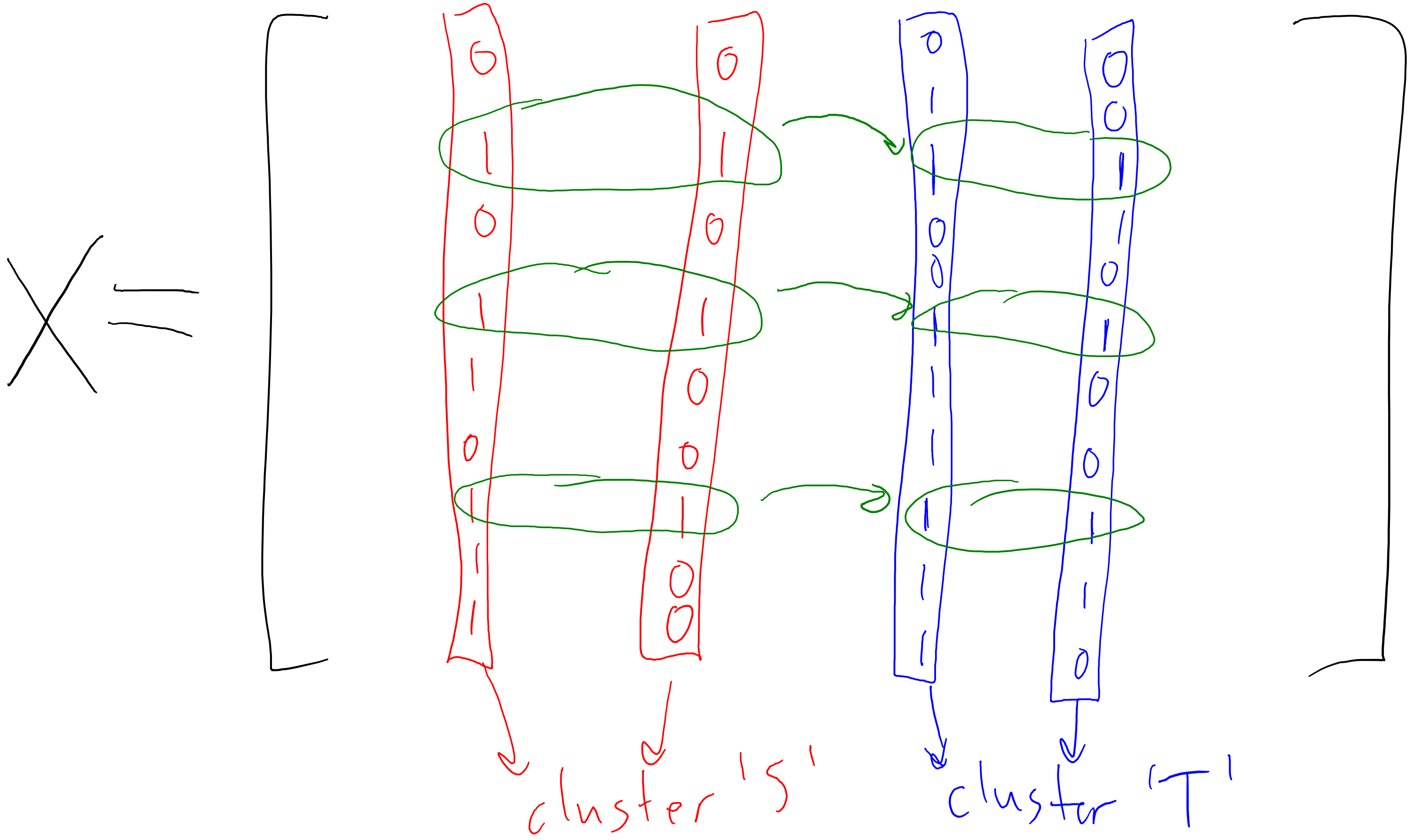
- We could cluster by columns:



- Apply clustering to X^T.

Association Rules

- Association rules (S => T): all ‘1’ in cluster S => all ‘1’ in cluster T.



Amazon Product Recommendation

- Amazon Product Recommendation works by columns:
  - Conceptually, you take the user-product matrix:

$$X = \begin{bmatrix} & & 1 \\ & & \\ & & \end{bmatrix}$$

user 'i' bought item 'j'.

  - And transpose it to make a product-user matrix:

$$X^T = \begin{bmatrix} & & \\ & 1 & \\ & & \end{bmatrix}$$

product 'i' was bought by user 'j'.

  - Find similar products as nearest neighbours among products.
    - Cosine similarity used to judge how 'close'

Supervised Learning Round 2: Regression

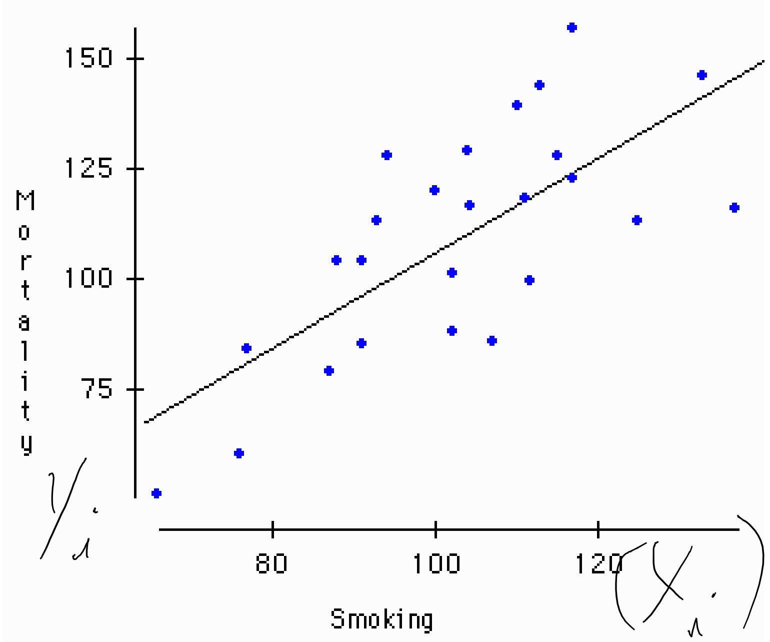
- We’re going to revisit supervised learning:

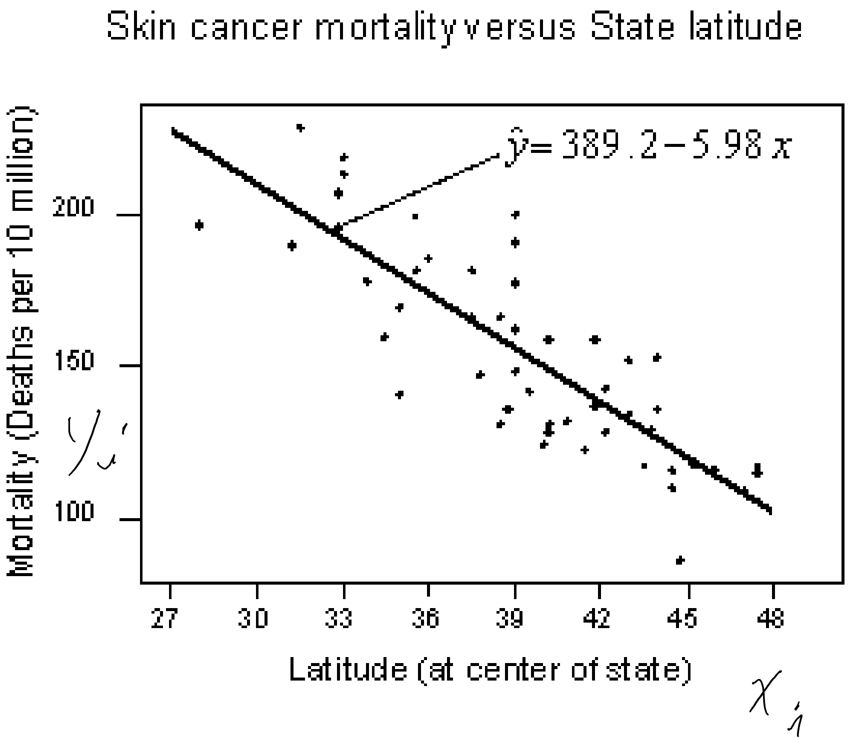
$$X = \begin{bmatrix} \quad\quad\quad\quad\quad\quad\quad\quad \end{bmatrix} \qquad y = \begin{bmatrix} \quad \end{bmatrix}$$

- Previously, we assumed y_i was discrete:
  - For example, y_i = ‘spam’ or y_i = ‘not spam’.
  - ‘Classification’.
- How we do we handle a continuous y_i?
  - For example, y_i = 10.34 cm.
  - ‘Regression’.

Example: Dependent vs. Explanatory Variables

- We want to discover relationship between factor and mortality:
  - Does number of lung cancer deaths change with number of cigarettes?
  - Does number of skin cancer deaths change with latitude?

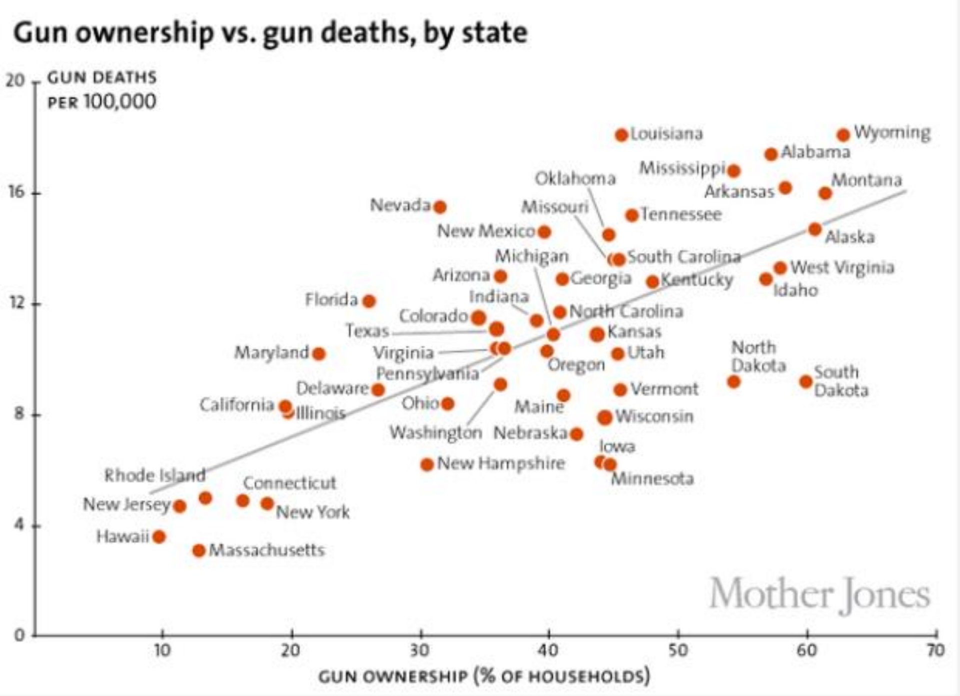




http://www.cvgs.k12.va.us:81/digstats/main/inferant/d_regrs.html
https://onlinecourses.science.psu.edu/stat501/node/11

Example: Dependent vs. Explanatory Variables

- We want to discover relationship between factor and mortality:
  - Does number of lung cancer deaths change with number of cigarettes?
  - Does number of skin cancer deaths change with latitude?
  - Does number of gun deaths change with gun ownership?



http://www.vox.com/2015/10/3/9444417/gun-violence-united-states-america

Handling Continuous Target Label

- One way to handle continuous y_i: discretize.
  - E.g., for 'age' could use {'age ≤ 20', '20 < age ≤ 30', 'age > 30'}.
  - Now can apply methods for classification to do regression.
  - But coarse discretization loses resolution.
  - And fine discretization requires lots of data.
- We can adapt classification methods to perform regression.
  - Next time: regression trees, generative models, non-parametric models.
- Today: one of oldest, but still most popular/important methods:
  - Linear regression based on squared error.

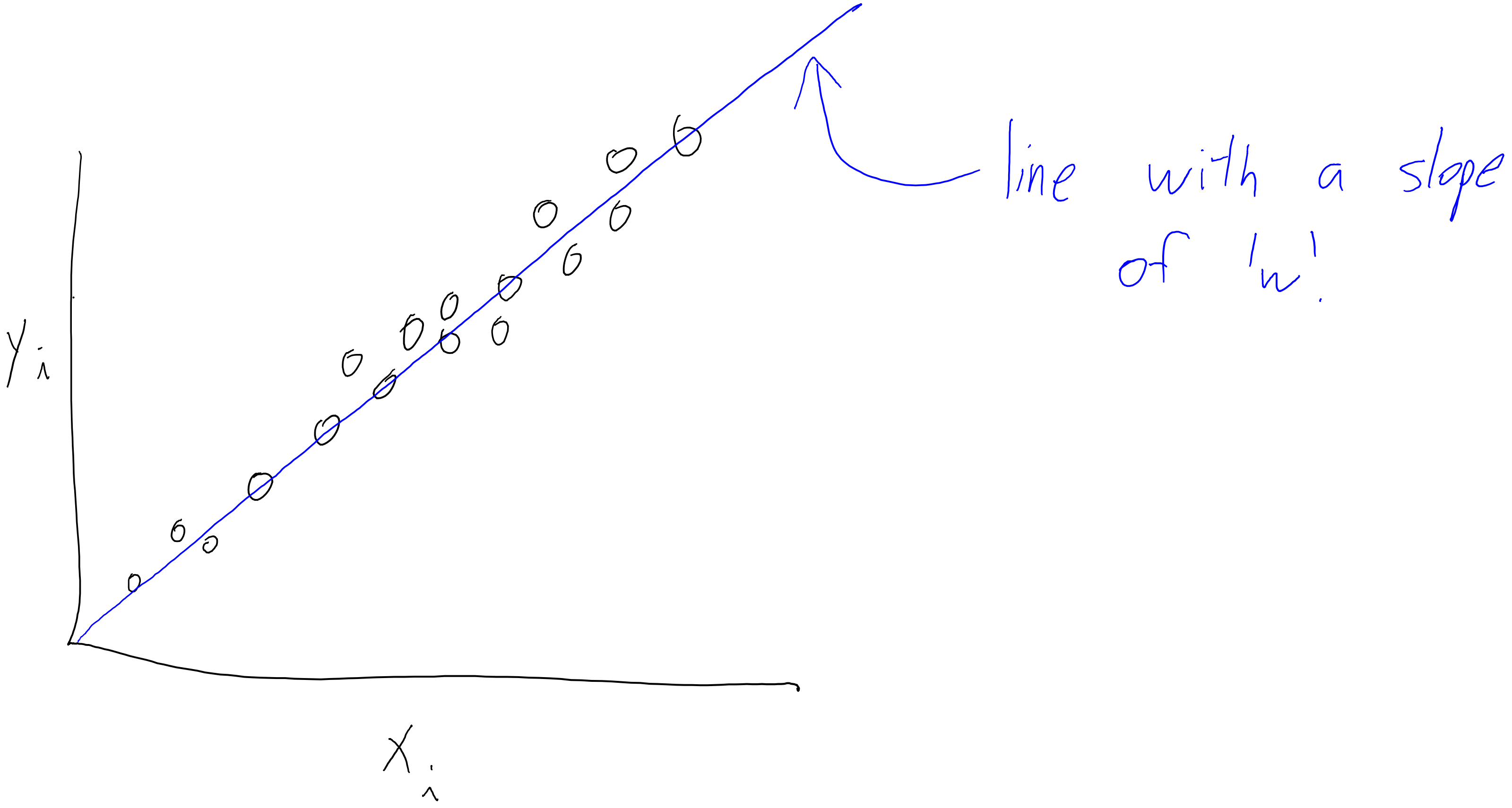
Linear Regression in 1 Dimension

- Assume we only have 1 feature:
  - For example, x_i is number of cigarettes, y_i is number of lung cancer deaths.
- Linear regression models y_i is a linear function of x_i:

$$y_i = w x_i$$

- The parameter 'w' is the weight or regression coefficient of x_i.
- As x_i changes, slope 'w' affects the rate that y_i increases/decreases:
  - Positive 'w': y_i increase as x_i increases.
  - Negative 'w': y_i decreases as x_i increases.

Linear Regression in 1 Dimension



Least Squares Objective

- Our linear model:

$$y_i = w x_i$$

- Classic way to set slope 'w' is minimizing sum of squared errors:

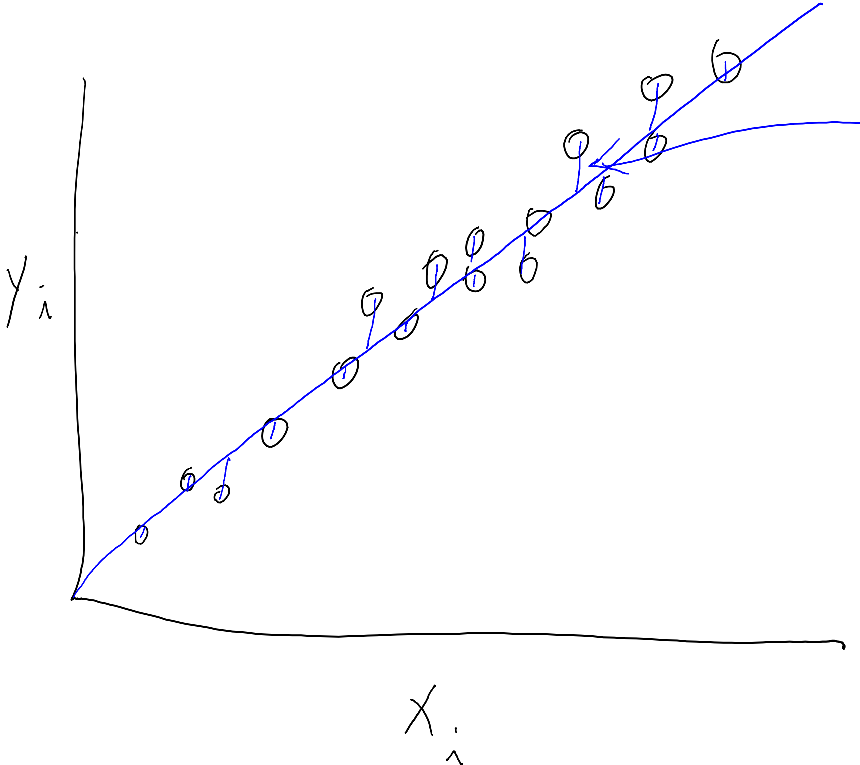
$$\underset{w \in \mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - w x_i)^2$$

- There are some justifications for this choice.
  - Assuming errors are Gaussian or using 'central limit theorem'.
- But usually, it is done because it is easy to compute.

Least Squares Objective

- Classic way to set slope 'w' is minimizing sum of squared errors:

$$\underset{w \in \mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - w x_i)^2$$

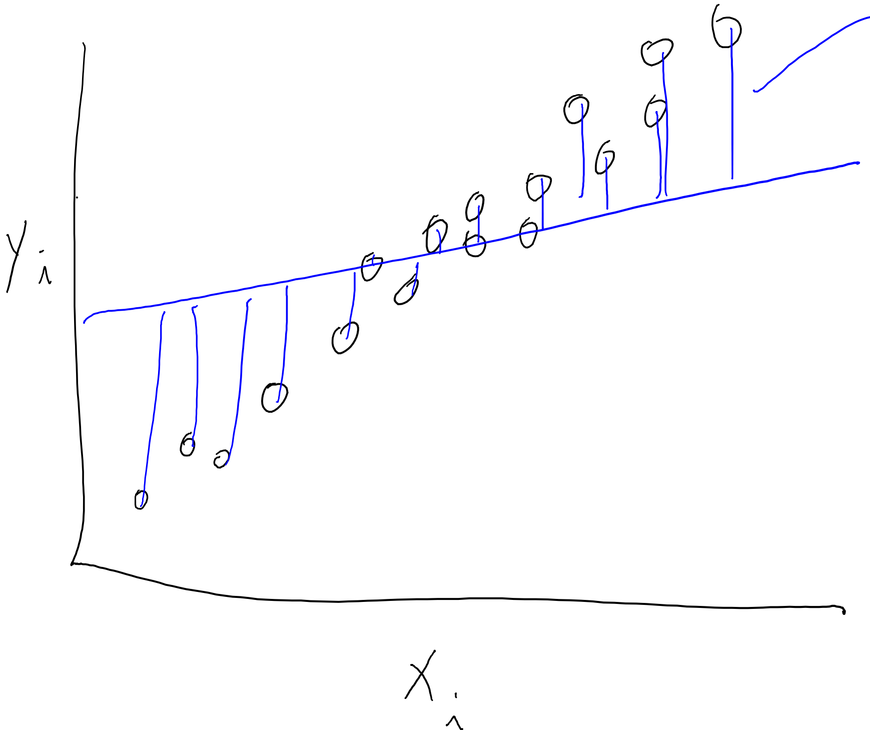


"errors" $(y_i - w x_i)$,
if these are small
then the line predicts
y_i accurately.

Least Squares Objective

- Classic way to set slope ‘w’ is minimizing sum of squared errors:

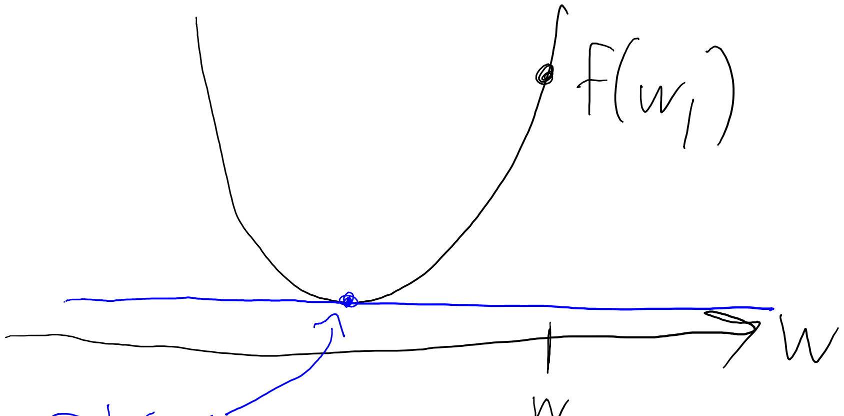
$$\underset{w \in \mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - w x_i)^2$$



If these vertical distances, $(y_i - w x_i)$, are large, then the line does a bad job of predicting y_i.

Minimizing a Differential Function

- Derivative-based approach to minimizing differentiable function ‘f’:
  1. Take the derivative of ‘f’.
  2. Find points ‘w’ where the derivative is equal to 0.
  3. Take the value among these points with the smallest f(w).
     (This assumes minimizer exists, if not sure then check that f’’(w) > 0.)



$f'(w)=0$, $f''(w)>0$, so this is a local minimizer.

Least Squares Objective

- Solving for 'w' that minimizes sum of squared errors:

Let $f(w) = \frac{1}{2}\sum_{i=1}^{n}(y_i - wx_i)^2 = \frac{1}{2}(y_1 - wx_1)^2 + \frac{1}{2}(y_2 - wx_2)^2 + \dots + \frac{1}{2}(y_n - wx_n)^2$

Then $f'(w) = -\sum_{i=1}^{n}(y_i - wx_i)x_i = -(y_1 - wx_1)x_1 - (y_2 - wx_2)x_2 - \dots - (y_n - wx_n)x_n$

If $f'(w) = 0$, then $-\sum_{i=1}^{n}[y_i x_i] + \sum_{i=1}^{n}[wx_i^2] = 0$, or

$$\sum_{i=1}^{n} y_i x_i = w\left[\sum_{i=1}^{n} x_i^2\right].$$

This gives $\boxed{w = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}}$

Least Squares Objective

- Checking that this is minimum:

$$f'(w) = -\sum_{i=1}^{n} (y_i - wx_i)x_i, \quad \text{so}$$

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

We know that (any real number)^2 cannot be negative, so $\sum_{i=1}^{n} x_i^2 \geq 0$. Thus, $f''(w) \geq 0$ for any 'w', and any w where $f'(w) = 0$ is a <u>minimizer</u>.

Motivation: Combining Explanatory Variables

- Smoking is not the only contributor to lung cancer.
  - For example, environmental factors like exposure to asbestos.
- How can we model the combined effect of smoking and asbestos?
- We can do this with a higher-dimensional linear function:

$$y_i = w_1 x_{i1} + w_2 x_{i2}$$

- Now we have a weight w_j for each feature 'j'.
- If we have 'd' features, the d-dimensional linear model is:

$$y_i = w_1 x_{i1} + w_2 x_{i2} + \dots + w_d x_{id}$$

Least Squares in d-Dimensions

- The 'd'-dimensional linear model:

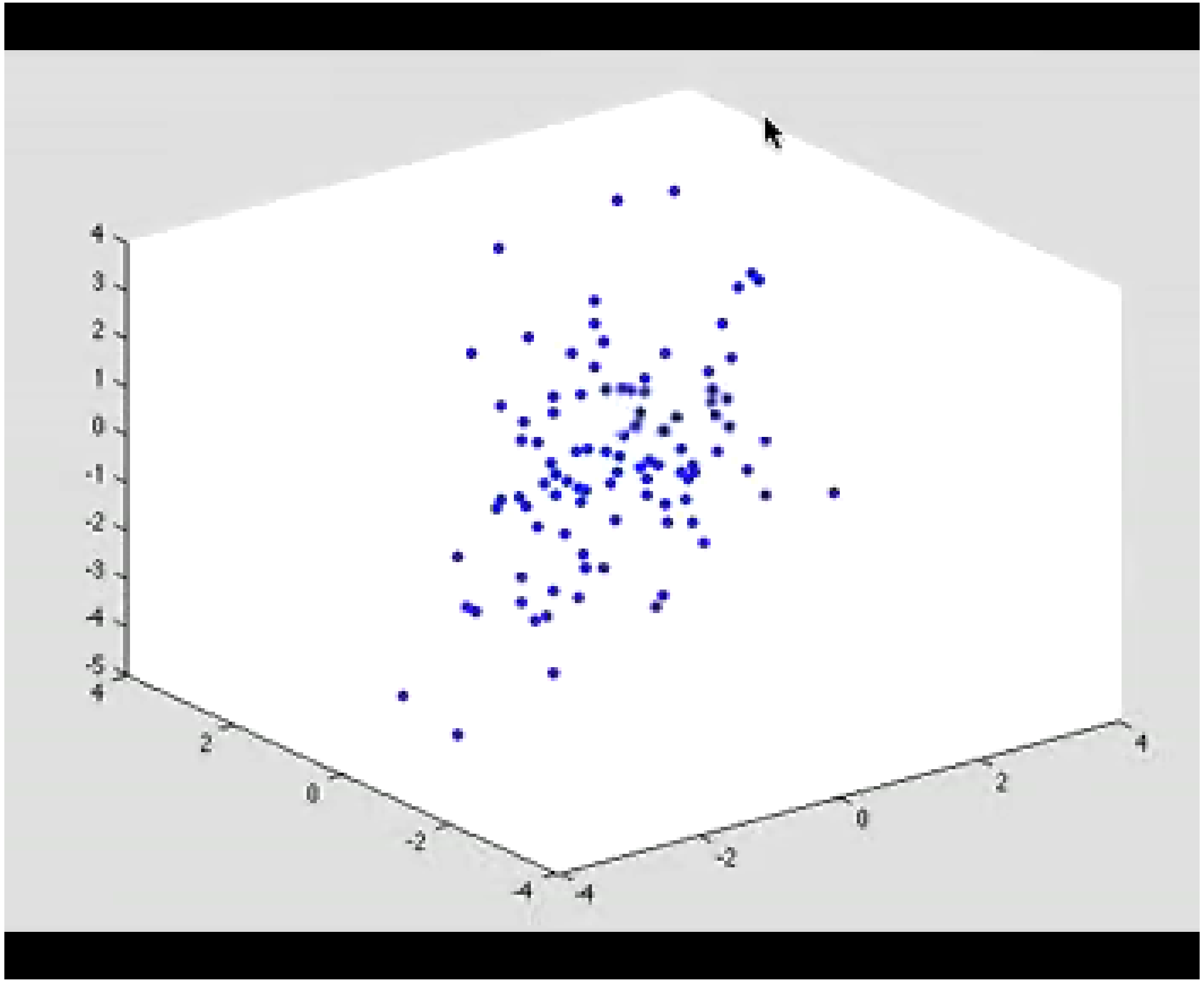
$$y_i = w_1 x_{i1} + w_2 x_{i2} + \cdots + w_d x_{id}$$
$$= \sum_{j=1}^{d} w_j x_{ij} \boxed{= w^T x_i} \rightarrow \text{the "inner product" between } w \text{ and } x_i.$$

- The general linear least squares model:

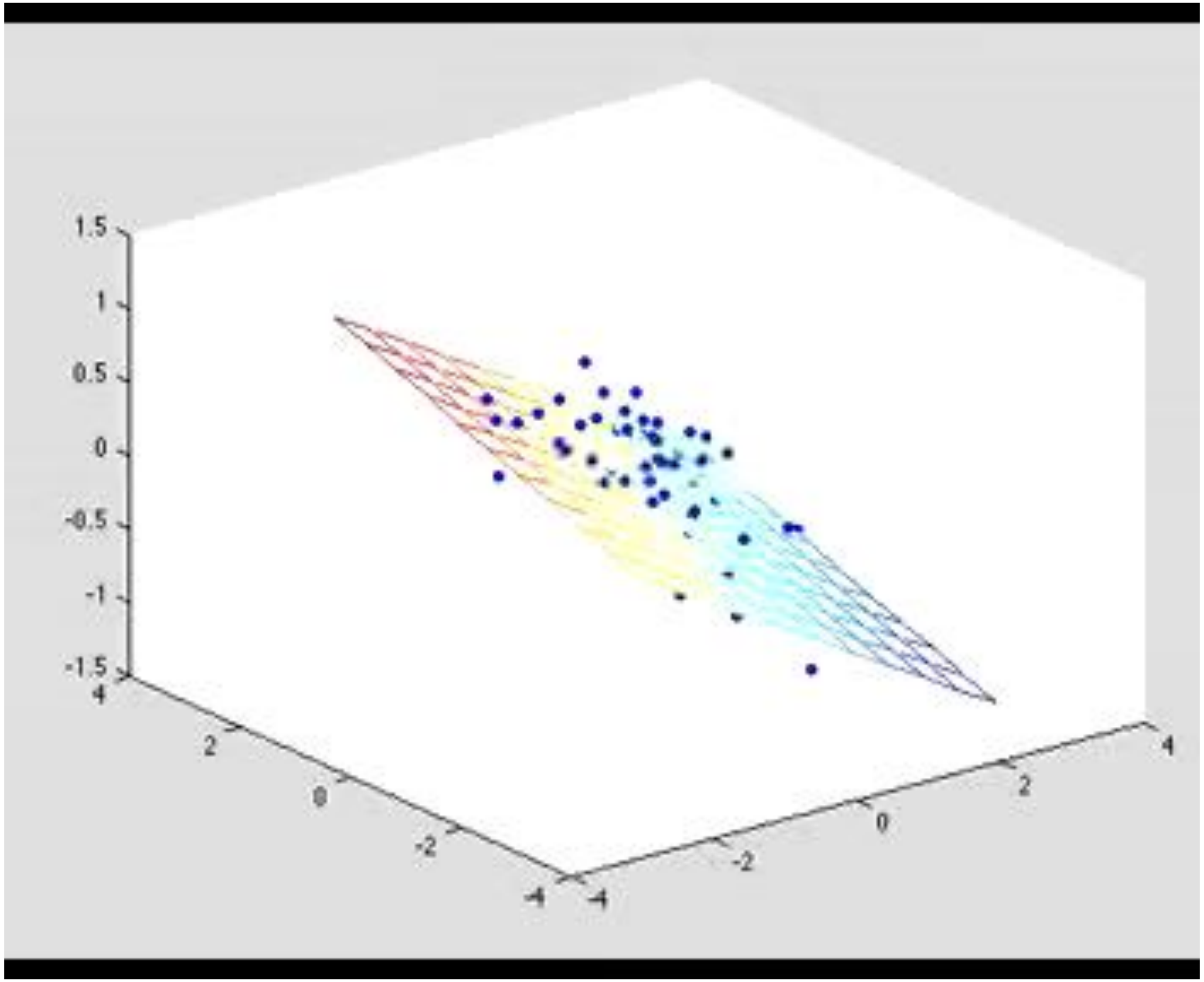
$$\underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{n} (y_i - w^T x_i)^2$$

- This is different than fitting each w_j individually.

Least Squares in 2-Dimensions



Least Squares in 2-Dimensions



Partial Derivatives and Gradient Vector

- Consider a multivariate real-valued function 'f'.

For example, $f(w_1, w_2) = 3w_1 + w_2^2 + w_1 w_2 + c.$

- Partial derivative with respect to 'j':
  - Derivative if we treat all other variables as fixed constants.

In example: $\frac{\partial f}{\partial w_1} = 3 + 0 + w_2 + 0 = 3 + w_2$

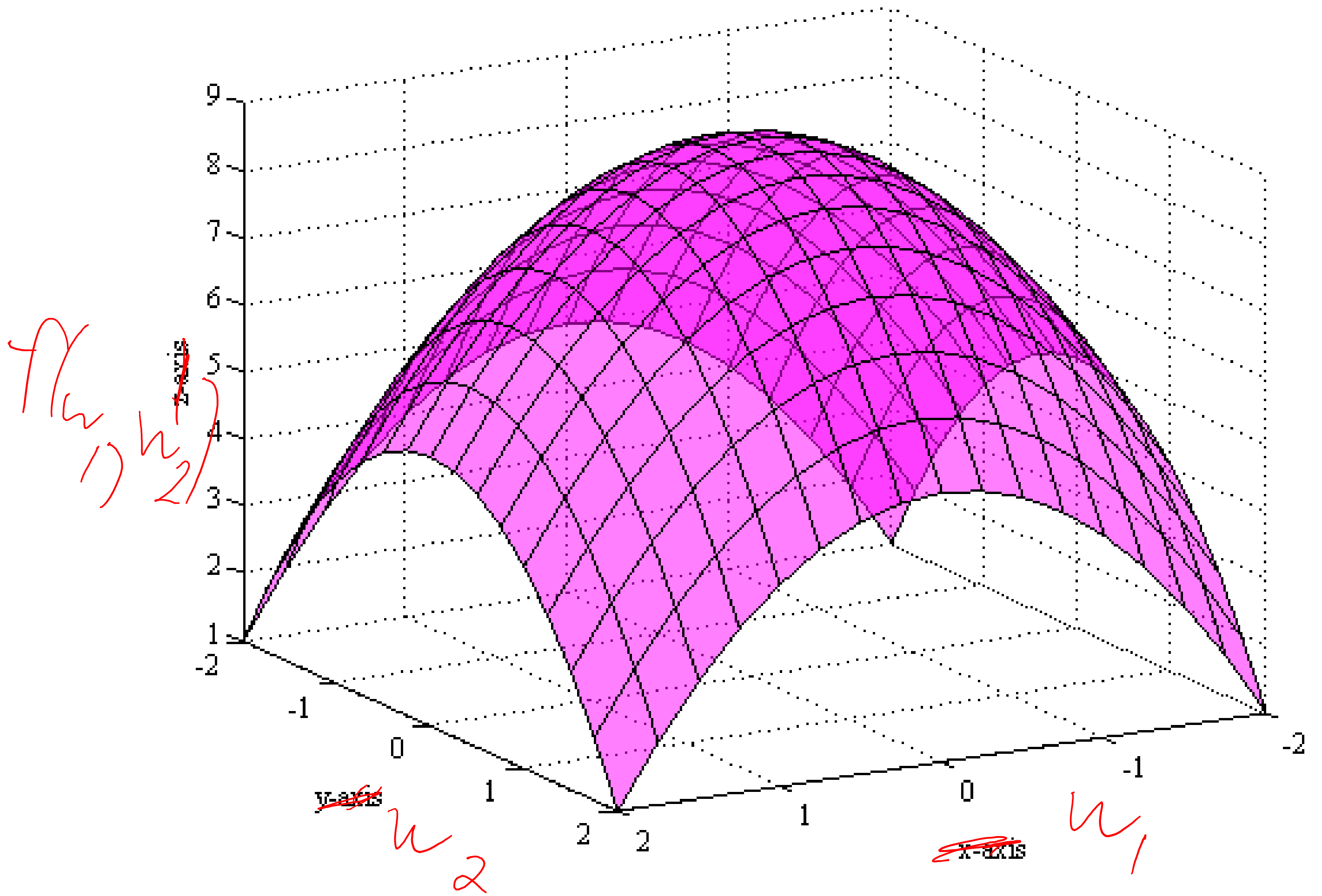
$\frac{\partial f}{\partial w_2} = 0 + 2w_2 + w_1 + 0 = 2w_2 + w_1$

"nabla"

- Gradient is vector with partial derivative 'j' in position 'j':

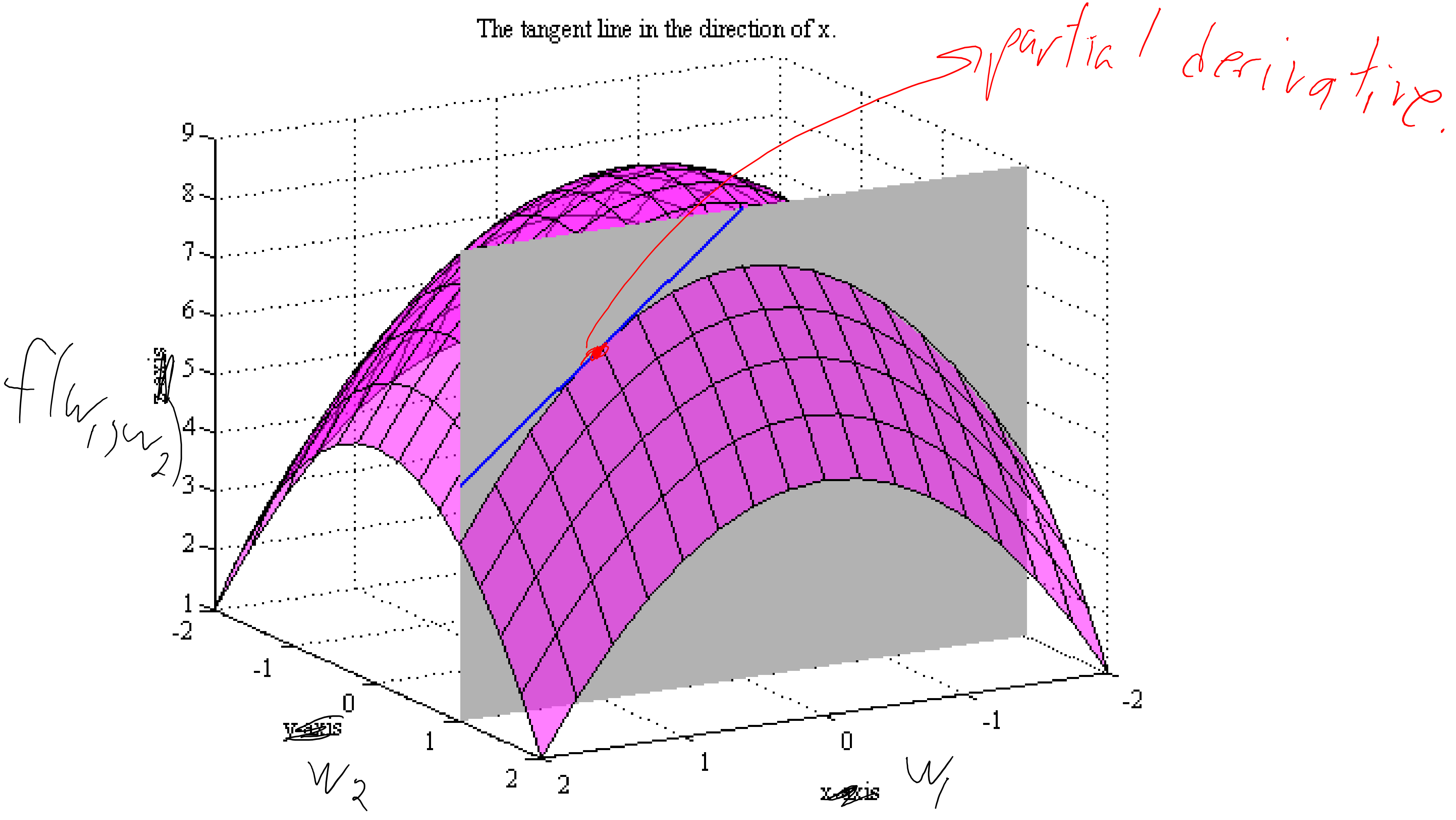
In example, $\nabla f(w_1, w_2) = \begin{bmatrix} \frac{\partial f}{\partial w_1} \\ \frac{\partial f}{\partial w_2} \end{bmatrix} = \begin{bmatrix} 3 + w_2 \\ 2w_2 + w_1 \end{bmatrix}$

Partial Derivatives and Gradient Vector



http://msemac.redwoods.edu/~darnold/math50c/matlab/pderiv/index.xhtml

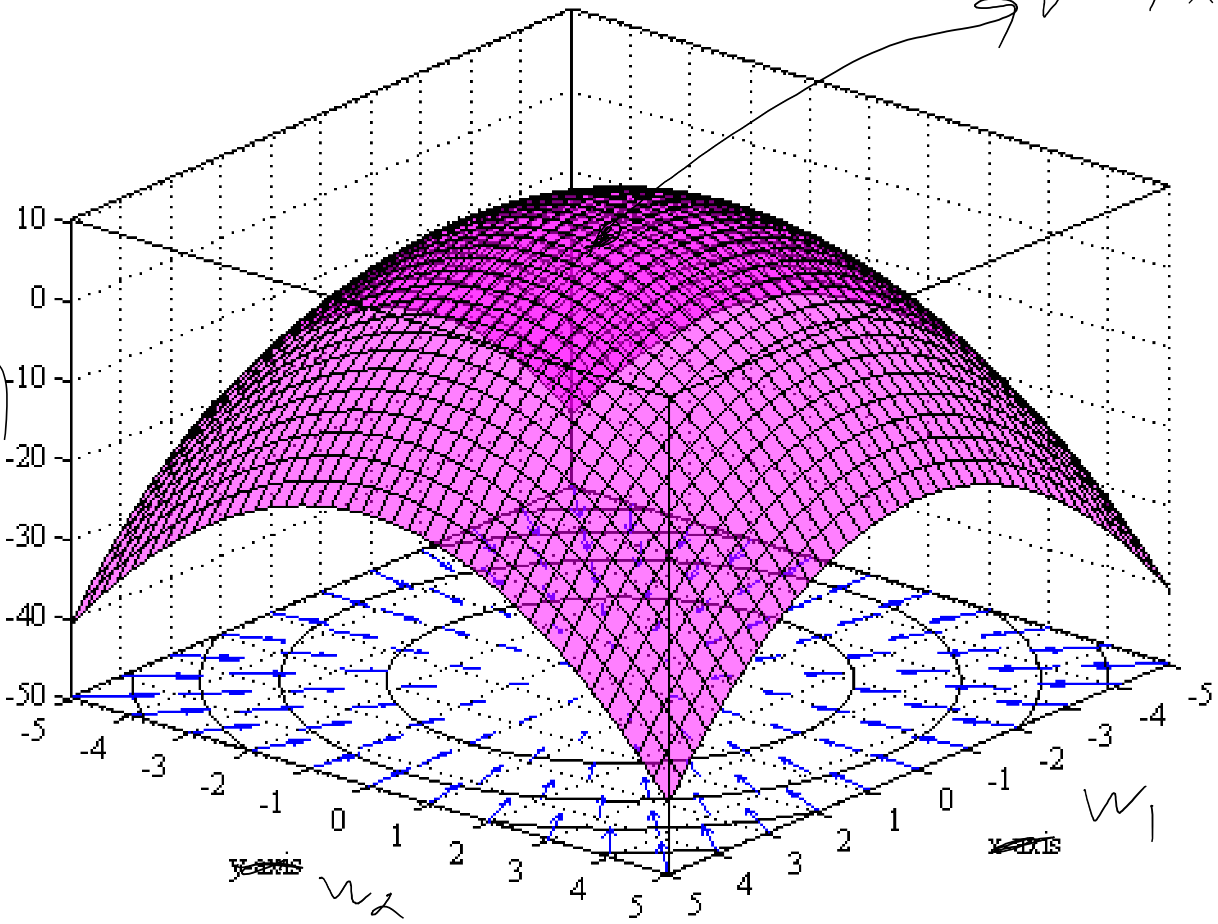
Partial Derivatives and Gradient Vector



Partial Derivatives and Gradient Vector

maximum: $\nabla f(w_1, w_2) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

$f(w_1, w_2)$

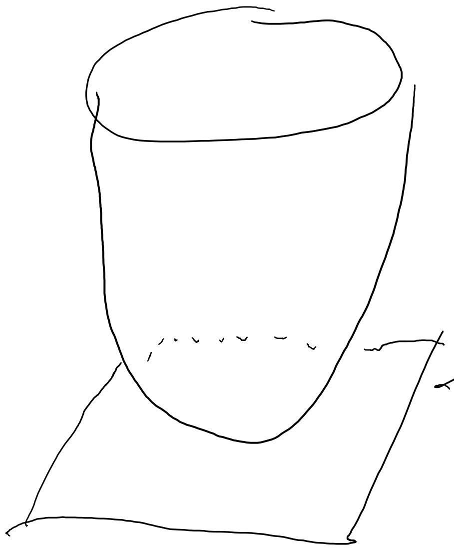


http://msemac.redwoods.edu/~darnold/math50c/matlab/pderiv/index.xhtml

Minima of Multivariate Functions

- To minimize a multivariate function (in principle):
  1. Find stationary points where $\nabla f(w) = 0$ (generalizes of $f'(w) = 0$).
  2. Take the value among these points with smallest $f(w)$.

  (This again assumes minimizer exists. If not sure, need to check that 'Hessian' matrix $\nabla^2 f(w)$ of second partial derivatives has non-negative eigenvalues.)



at minimizer, gradient is (0,0).

Least Squares Gradient

Objective is $f(w) = \frac{1}{2}\sum_{i=1}^{n}(y_i - w^T x_i)^2$.

Partial derivative with respect to w_j:

$$\frac{\partial f}{\partial w_j} = -\sum_{i=1}^{n}(y_i - w^T x_i)x_{ij}.$$

Gradient vector:

$$\nabla f(w) = \begin{bmatrix} -\sum_{i=1}^{n}(y_i - w^T x_i)x_{i1} \\ -\sum_{i=1}^{n}(y_i - w^T x_i)x_{i2} \\ \vdots \\ -\sum_{i=1}^{n}(y_i - w^T x_i)x_{id} \end{bmatrix}$$

Summary

- Regression considers the case of a continuous y_i.
- Least squares is a classic method for fitting linear models.
- Differentiation leads to a closed-form solution for slope 'w'.
- Gradient is vector containing partial derivatives wrt all variables.

- Next time:
  - Non-linear regression methods.