Machine Learning

CS 165B

Prof. Matthew Turk

Monday, April 25, 2016

Today

- Linear learning models (Ch 7)

Notes

- GauchoSpace expiration notifications – FIXED
  - I extended all of them until late June
- Midterm – Monday, May 2, in class
  - Covers material through Wednesday
  - Brief review in Wednesday's lecture
  - Practice midterm will be supplied by this weekend
  - Closed book/notes
    - Exception: You may bring one 8.5"x11" sheet of paper with your notes (both sides)
    - I'll also provide some information, formulas, etc. (will be included with the practice midterm)
- HW#3 will be posted on Friday

Notes

- NO CLASS MEETING THIS WEDNESDAY
  - Instead, I will post an “audio lecture” – PowerPoint with audio
  - Use the regular class time (or soon thereafter) to listen to this lecture on your own (or with a group)

- NO OFFICE HOURS TOMORROW for me
  - Instead, I’ll hold office hours 9:30-11:10am on Thursday
  - A good opportunity to ask questions about Wednesday’s audio lecture

Linear Learning Models

Chapter 7 in the textbook

And SVMs, kernel methods, perceptrons…

Key statistical concepts

- Mean – average; expected value of a variable

  $\mu_x = E[X] = \sum_{i=1}^{n} x_i p_i$ or $\int x\, p(x)\, dx$

- Variance – a measure of the spread of a variable

  $Var(X) = \sigma_x{}^2 = E[(X - \mu_x)^2] = E[X^2] - \mu_x{}^2$

  Standard deviation: $\sigma_x = \mathrm{Sqrt}(\sigma_x{}^2)$

- Estimating mean and variance from data $\{x_\mathrm{i}\}$

  Sample mean: $\hat{\mu}_x = \frac{1}{n}\sum_i x_i$

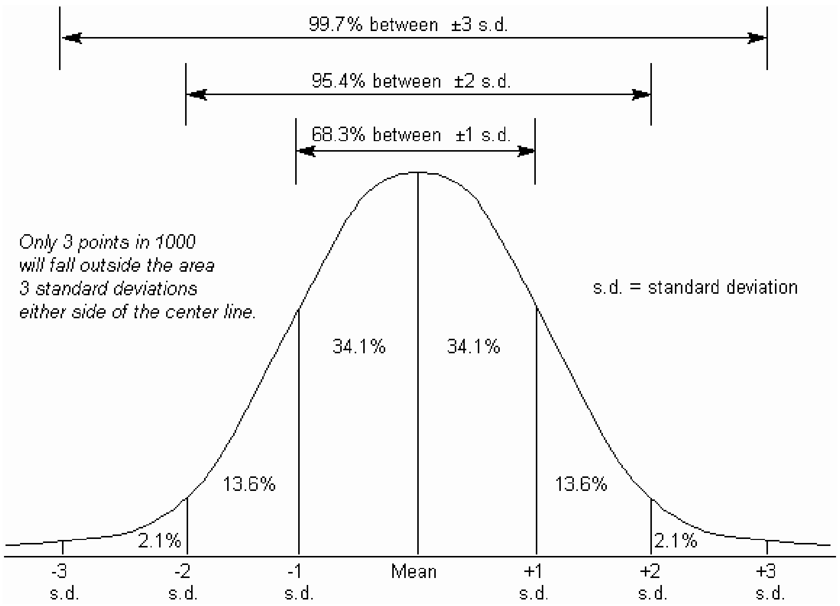
  Sample variance: $\hat{\sigma}_x{}^2 = \frac{1}{n}\sum_i (x_i - \hat{\mu}_x)^2$ or $s = \frac{1}{n-1}\sum_i (x_i - \hat{\mu}_x)^2$

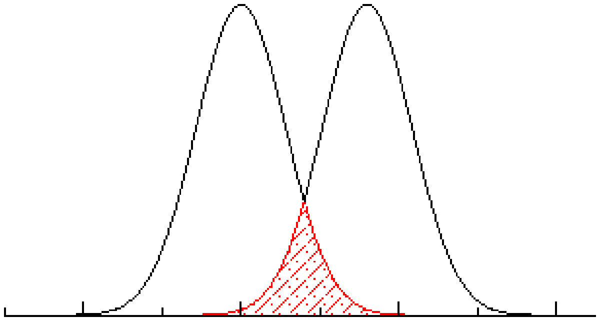
- Covariance – a measure of how two variables change together

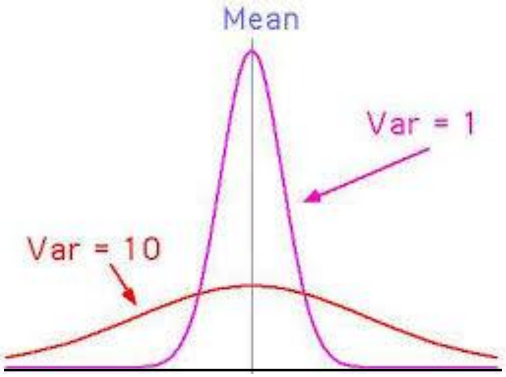
  $Cov(X, Y) = \sigma_{xy} = E[(X - \mu_x)(Y - \mu_y)] = E[XY] - \mu_x\, \mu_y$

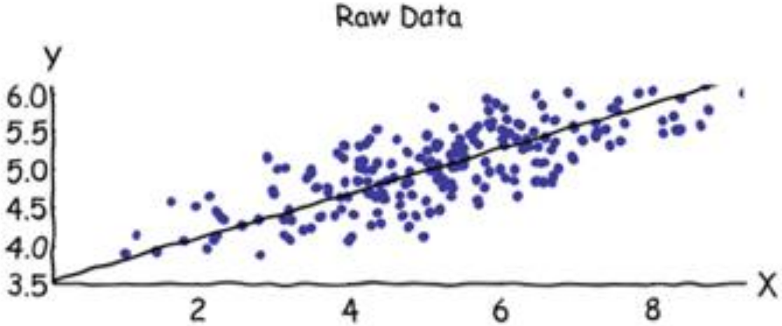
  Sample covariance: $\hat{\sigma}_{xy} = \frac{1}{n}\sum_i (x_i - \hat{\mu}_x)(y_i - \hat{\mu}_y)$ or $\frac{1}{n-1}\sum_i (x_i - \hat{\mu}_x)(y_i - \hat{\mu}_y)$

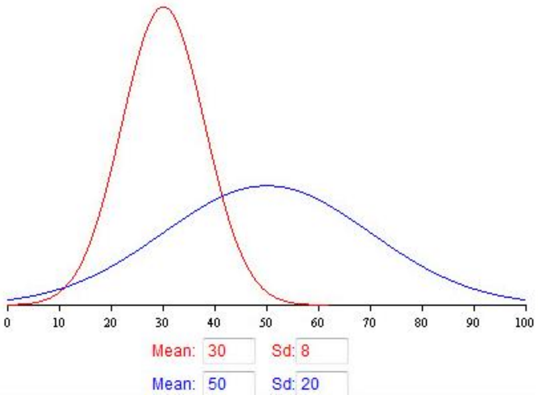
Key statistical concepts (cont.)











Key statistical concepts (cont.)

- Covariance matrix Σ
  - For n variables $\boldsymbol{X}$, an n x n matrix whose elements are $\mathrm{Cov}(X_i, X_j)$
  - Diagonal entries are variances: $Cov(X_i, Xi) = Var(Xi)$

Scatter matrix

Sample covariance: $\hat{\Sigma}_{ij} = \frac{1}{k}\sum_k (x_{ik} - \hat{\mu}_i)(x_{jk} - \hat{\mu}_j) = \frac{1}{k}S$

If X is a matrix that holds all the zero-centered samples as column vectors, then $$\hat{\Sigma} = \frac{1}{k}XX^T$$

- If variables x and y are uncorrelated, then

$$Cov(X, Y) = \sigma_{xy} = 0$$

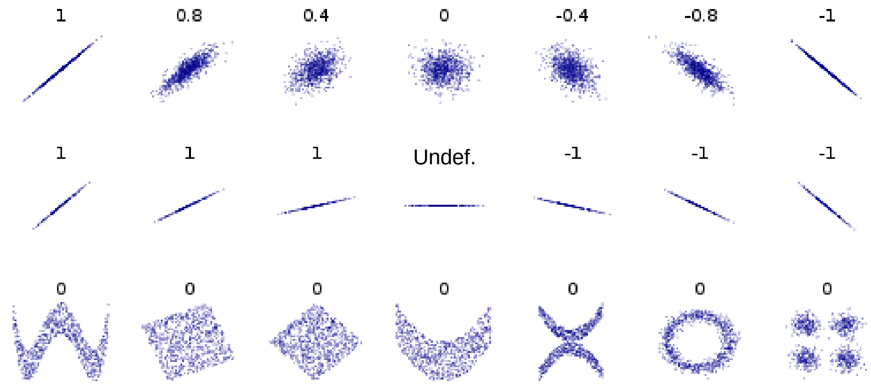
  - Uncorrelated variables: knowing the value of X (or Y) tells you nothing about the value of Y (or X)
  - So the covariance matrix for uncorrelated variables is a diagonal matrix consisting of the n variances

Examples

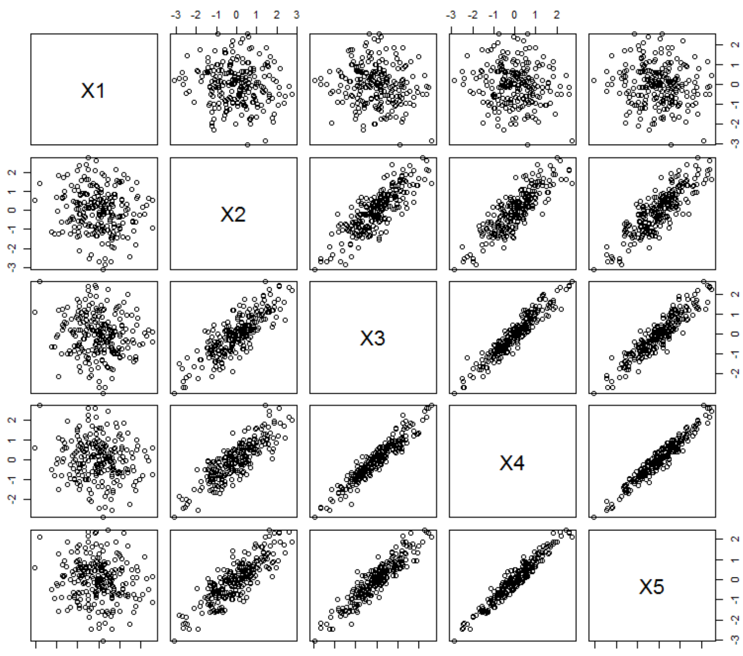
2D data and their correlation coefficient (ρ) values

$$\rho_{x,y} = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

$$0 \leq \rho \leq 1$$



Not a useful measure for nonlinear data!



Visualizing a 5-variable covariance matrix (symmetric about the diagonal)

Linear models

- Linear models are geometric models for which the regression functions or decision boundaries are linear
  - Lines, planes, hyperplanes (N-dimensional planes)
- Definition of a linear function:

$$y = f(ax_1 + bx_2) = af(x_1) + bf(x_2)$$

or in matrix notation, a linear transformation:

$$\boldsymbol{y} = M\boldsymbol{x}$$

- An affine function is a linear function plus a constant

$$f_{\text{aff}}(x) = f_{\text{lin}}(x) + c$$

In matrix notation:

$$\boldsymbol{y} = M\boldsymbol{x} + \boldsymbol{c}$$

Using homogeneous coordinates:

$$\boldsymbol{y} = \boldsymbol{M}'\boldsymbol{x_h}$$

$$\boldsymbol{y} = M\boldsymbol{x} + \boldsymbol{c}$$

$$\boldsymbol{y} = [\boldsymbol{M} \quad \boldsymbol{c}] \begin{bmatrix} \boldsymbol{x} \\ 1 \end{bmatrix}$$

$$\boldsymbol{y} = M'\boldsymbol{x_h}$$

$$\boldsymbol{y} = M\boldsymbol{x} + \boldsymbol{c}$$

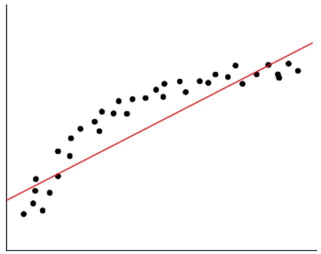
$$\begin{bmatrix} \boldsymbol{y} \\ 1 \end{bmatrix} = \begin{bmatrix} M & \boldsymbol{c} \\ \boldsymbol{0} & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ 1 \end{bmatrix}$$

$$\boldsymbol{y_h} = M''\boldsymbol{x_h}$$

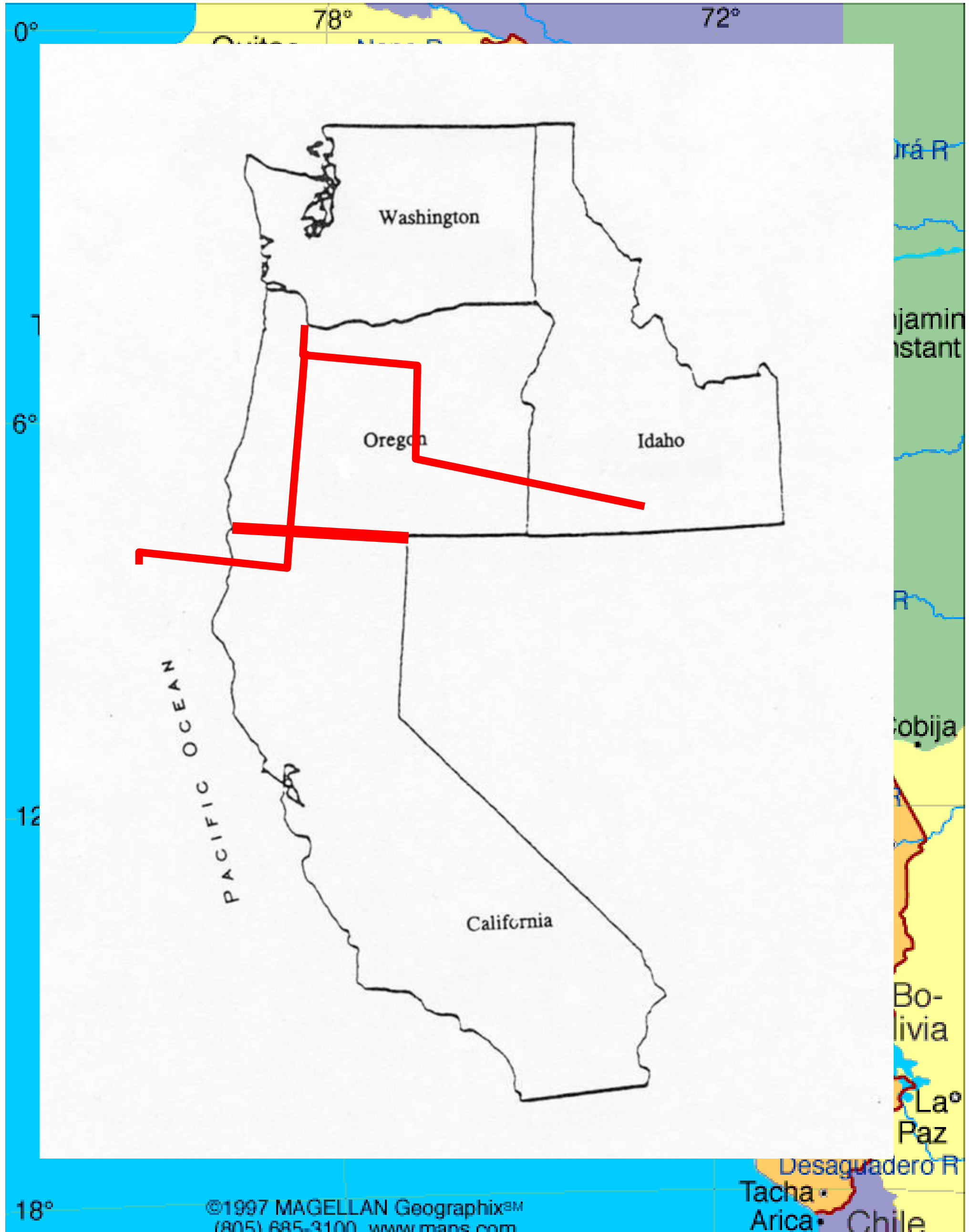
So we can use the term ***linear models*** to include ***affine models***

Linear models

- Linear learning models are widely used because
  - Many functions can be reasonably approximated as linear, or at least as piecewise linear
  - They're simple, and thus easy to train
  - The math is tractable
  - They avoid over-fitting – i.e., they generalize well when the data is very noisy
- However, they are prone to under-fitting
  - I.e., over-simplifying a more complicated function

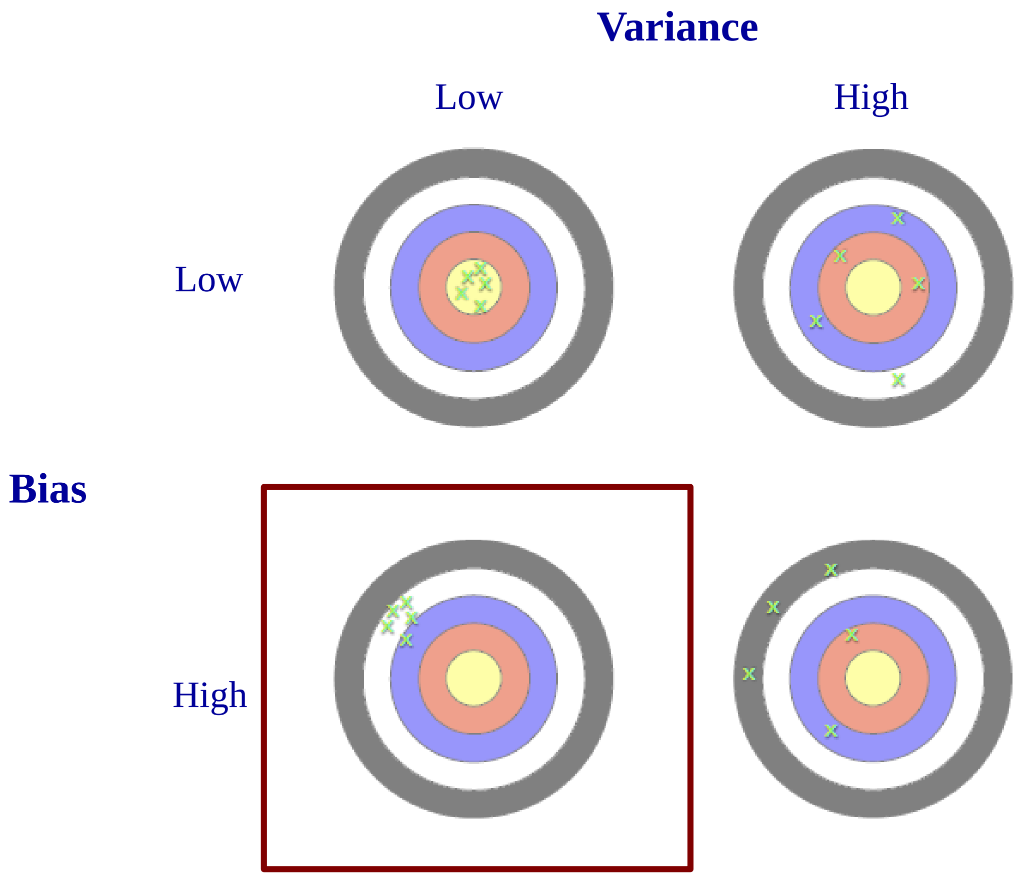


- For example, learning borders from sample data
  - The border between California and Oregon – linear
  - The border between Texas and New Mexico – piecewise linear
  - The border between Texas and Oklahoma – piecewise linear approx.
  - The border between Peru and Brazil – complicated!


0°
78°
72°
Washington
Idaho
California
PACIFIC OCEAN
6°
18°
Bo-
Paz
Desaguadero R
Tacha
Arica
Chile
©1997 MAGELLAN Geographix℠
(805) 685-3100 www.maps.com

Linear models

- Linear models tend to have low variance but high bias



Low variance – stability, robustness

Performance on different testing sets should be similar

High bias – limited accuracy, underfitting, systematic (but consistent) errors

Parametric models

- Linear models are parametric models
  - Within a given family of models (e.g., lines or planes), we just need to learn the model parameters (e.g., 2 or 3 coefficients)
- We'll also consider nonparametric models
  - No explicit assumption about the shape of the model
- For example, in a 2D classification problem we could use linear decision boundaries (lines) as a parametric model, or the nearest-neighbor approach (minimum distance) as a non-parametric model
- This distinction is also important in density estimation – estimating a probability distribution or density from data
  - E.g., in parametric estimation, we might assume the pdf is Gaussian, so the task becomes estimating the Gaussian parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Linear least-squares regression

- Regression learns a function (the regressor) that is a mapping $\hat{f} : \mathcal{X} \rightarrow \mathbb{R}$; it's learned from examples, $(x_i, f(x_i))$
  - I.e., the target variable (output $\hat{f}(x)$) is real-valued
- Linear regression – the function is linear
  - Fit a line/plane/hyperplane to the data
- The difference between f and $\hat{f}$ is known as the residual ϵ

$$\epsilon_i = f(x_i) - \hat{f}(x_i)$$

- The least squares method minimizes the sum of the squared residuals – i.e., find $\hat{f}$ that minimizes $\sum_i {\epsilon_i}^2$ on the training data
- Univariate or multivariate regression
  - Univariate – one input variable
  - Multivariate – multiple input variables

Linear least-squares regression example

- We wish to find the relationship between the height and weight of adults
  - Data: n measurements, $(h_i, w_i) \rightarrow (input, output)$
  - Parametric linear model: $w = a + bh \quad \Rightarrow \quad w_i = a + bh_i + \epsilon_i$
  - Residual: $\epsilon_i = w_i - (a + bh_i)$
  - Find (a, b) that minimizes $\sum_i [w_i - (a + bh_i)]^2$ on the training data
- To minimize, set the partial derivatives (wrt a and b) to zero and solve for a and b

$$\frac{\partial}{\partial a}\sum_{i=1}^{n}(w_i - (a + bh_i))^2 = -2\sum_{i=1}^{n}(w_i - (a + bh_i)) = 0 \qquad \Rightarrow \hat{a} = \overline{w} - \hat{b}\overline{h}$$

$$\frac{\partial}{\partial b}\sum_{i=1}^{n}(w_i - (a + bh_i))^2 = -2\sum_{i=1}^{n}(w_i - (a + bh_i))h_i = 0 \qquad \Rightarrow \hat{b} = \frac{\sum_{i=1}^{n}(h_i - \overline{h})(w_i - \overline{w})}{\sum_{i=1}^{n}(h_i - \overline{h})^2}$$

- So the regression model is $w = \hat{a} + \hat{b}h = \overline{w} + \hat{b}(h - \overline{h})$

Note that the regression line goes through $(\bar{h}, \bar{w})$

The regression coefficient

- The slope ($\hat{b}$) is the regression coefficient

$$\hat{b} = \frac{\sum_{i=1}^{n}(h_i - \overline{h})(w_i - \overline{w})}{\sum_{i=1}^{n}(h_i - \overline{h})^2} = \frac{n\sigma_{hw}}{n{\sigma_h}^2} = \frac{\sigma_{hw}}{{\sigma_h}^2}$$

- In general, the regression coefficient for a feature *x* and a target variable *y* is

covariance(*x*, *y*)

$$\hat{b} = \frac{\sigma_{xy}}{{\sigma_x}^2}$$

variance(*x*)

- We can simplify the problem by first normalizing the data
  - Find the data averages $(\bar{h}, \bar{w})$
  - Subtract the averages from the data: $h_i \leftarrow h_i - \bar{h}$

    $w_i \leftarrow w_i - \bar{w}$
  - This makes $\hat{a} = 0$, so we're just left with estimating the regression coefficient $\hat{b}$

Multivariate linear regression

- Most linear regression problems involve multiple input variables
  - E.g., estimate a patient's cholesterol level from several input variables
- In multivariate LR, there are N+1 regression parameters
- Linear regression equations:

x_{i0} = 1 (homogeneous notation)

Univariate

$$y_i = w_1 x_i + w_0 + \epsilon_i$$

Multivariate

$$y_i = w_2 x_{i2} + w_1 x_{i1} + w_0 x_{i0} + \epsilon_i$$

Column of 1s

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \end{bmatrix} \qquad \boldsymbol{X} = \begin{bmatrix} x_{12} & x_{11} & \boxed{\begin{matrix} x_{10} \\ x_{20} \\ \vdots \end{matrix}} \\ x_{22} & x_{21} & \\ \vdots & \vdots & \end{bmatrix} \qquad \boldsymbol{w} = \begin{bmatrix} w_2 \\ w_1 \\ w_0 \end{bmatrix} \qquad \boldsymbol{\epsilon_i} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \end{bmatrix}$$

Labels — Data (homogeneous) — Regression parameters — Residuals

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\epsilon}$$

Multivariate least-squares in matrix form

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\epsilon}$$

$$\widehat{\boldsymbol{w}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Least-squares solution $\widehat{\boldsymbol{w}}$

$$= \boldsymbol{S}^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Scatter matrix for $\boldsymbol{X}$

$$\boldsymbol{S} = \boldsymbol{X}^T\boldsymbol{X}$$

Linear regression function

$$y(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{w}$$

Using homogeneous coordinates

Note: Often $\boldsymbol{X}$ is written transposed from how it's defined here, so

$$\boldsymbol{y} = \boldsymbol{X}^T\boldsymbol{w} + \boldsymbol{\epsilon}$$

$$\widehat{\boldsymbol{w}} = (\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}\boldsymbol{y}$$

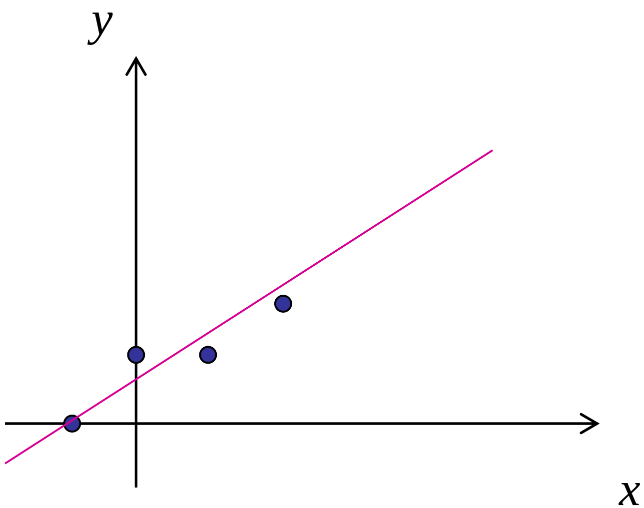
$$\boldsymbol{S} = \boldsymbol{X}\boldsymbol{X}^T$$

Need to understand in context

Simple linear regression example

Training set:
(-1, 0)
(0, 1)
(1, 1)
(2, 2)

inputs (x) outputs (y)



$$\hat{\boldsymbol{w}} = \begin{bmatrix} 0.6 \\ 0.7 \end{bmatrix} = \begin{bmatrix} \text{slope} \\ \text{y−intercept} \end{bmatrix}$$

Learn the regression function $y(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{w} = wx - t$

$$\boldsymbol{y} = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 2 \end{bmatrix} \qquad \boldsymbol{X} = \begin{bmatrix} -1 & 1 \\ 0 & 1 \\ 1 & 1 \\ 2 & 1 \end{bmatrix} \qquad \boldsymbol{w} = \begin{bmatrix} w \\ -t \end{bmatrix}$$

Homogeneous representation

$$\hat{\boldsymbol{w}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$$

$$= \left(\begin{bmatrix} -1 & 0 & 1 & 2 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} -1 & 1 \\ 0 & 1 \\ 1 & 1 \\ 2 & 1 \end{bmatrix} \right)^{-1} \boldsymbol{X}^T \boldsymbol{y} = \begin{bmatrix} 6 & 2 \\ 2 & 4 \end{bmatrix}^{-1} \boldsymbol{X}^T \boldsymbol{y}$$

$$= \begin{bmatrix} 0.2 & -0.1 \\ -0.1 & 0.3 \end{bmatrix} \begin{bmatrix} -1 & 0 & 1 & 2 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 0.6 \\ 0.7 \end{bmatrix}$$

$$y(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{w} = \begin{bmatrix} x & 1 \end{bmatrix} \begin{bmatrix} 0.6 \\ 0.7 \end{bmatrix} = 0.6x + 0.7$$

Feature correlation

- If the features in a multivariate regression problem with *d* input features are uncorrelated ($\sigma_{x_i x_j} = 0$ if $i \neq j$) then the problem reduces to *d* univariate problems
  – This relates to the task of feature construction – construct uncorrelated features to simplify the problem!
  – We may come back to this in Chapter 10 on features

Regularization

In a typical polynomial regression, $r(\boldsymbol{w}) = \|\boldsymbol{w}\|^2$ to discourage large coefficients

- Another way to formulate the multivariate least-squares problem is

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\epsilon}$$

$$\boldsymbol{w}^* = \underset{\boldsymbol{w}}{\operatorname{argmin}}\,(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})$$

(least squares minimization)

- Sometimes we'd like to provide constraints on the solution in order to avoid overfitting to the data
  - E.g., if we think the training data may not be representative, or we have external knowledge about the problem beyond the data

- One way to do this is through regularization

$$\boldsymbol{w}^* = \underset{\boldsymbol{w}}{\operatorname{argmin}}\,(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) + \lambda\,\underbrace{r(\boldsymbol{w})}_{\text{Regularization function}}$$

λ is a scalar determining the amount of regularization

  - So now when we optimize (minimize) to choose $\boldsymbol{w}^*$, λ is involved