Machine Learning

Regression

Indian Institute of Information Technology
Sri City, Chittoor

Today's Agenda

- Introduction to Regression
- Introduction to Linear Regression
- Additional Resources

Introduction to Regression

- Regression analysis is the part of statistics that investigates the relationship between two or more variables related in a nondeterministic fashion.
- Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data.
- One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable.
- For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.
- Before attempting to fit a linear model to observed data, a modeler should first determine whether or not there is a relationship between the variables of interest.

Introduction to Linear Regression

- This does not necessarily imply that one variable causes the other (for example, higher SAT scores do not cause higher college grades), but that there is some significant association between the two variables

- A scatterplot can be a helpful tool in determining the strength of the relationship between two variables.

- If there appears to be no association between the proposed explanatory and dependent variables (i.e., the scatterplot does not indicate any increasing or decreasing trends), then fitting a linear regression model to the data probably will not provide a useful model.

Introduction to Linear Regression

- The simplest deterministic mathematical relationship between two variables x and y is a linear relationship.

$$y = b_0 + b_1 x$$

- The set of pairs (x, y) for which determines a straight line with slope b_1 and y-intercept b_0.

- More generally, the denoted by x and will be called the independent, predictor, or explanatory variable.

- For fixed x, the second variable will be random; we denote this random variable and its observed value by Y and y, respectively, and refer to it as the dependent or response variable.

Introduction to Linear Regression

- Usually observations will be made for a number of settings of the independent variable.
- Let $x_1, x_2, \ldots, x_n$ denote values of the independent variable for which observations are made, and let Y_i and y_i , respectively, denote the random variable and observed value associated with.
- The available bivariate data then consists of the n pairs

$$(x_1, y_1), (x_2, y_2), \ldots\ldots, (x_n, y_n).$$

- A picture of this data called a **scatter plot** gives preliminary impressions about the nature of any relationship.
- In such a plot, each (x_i, y_i) is represented as a point plotted on a two-dimensional coordinate system.

Scatter Plot Example: Linear Regression

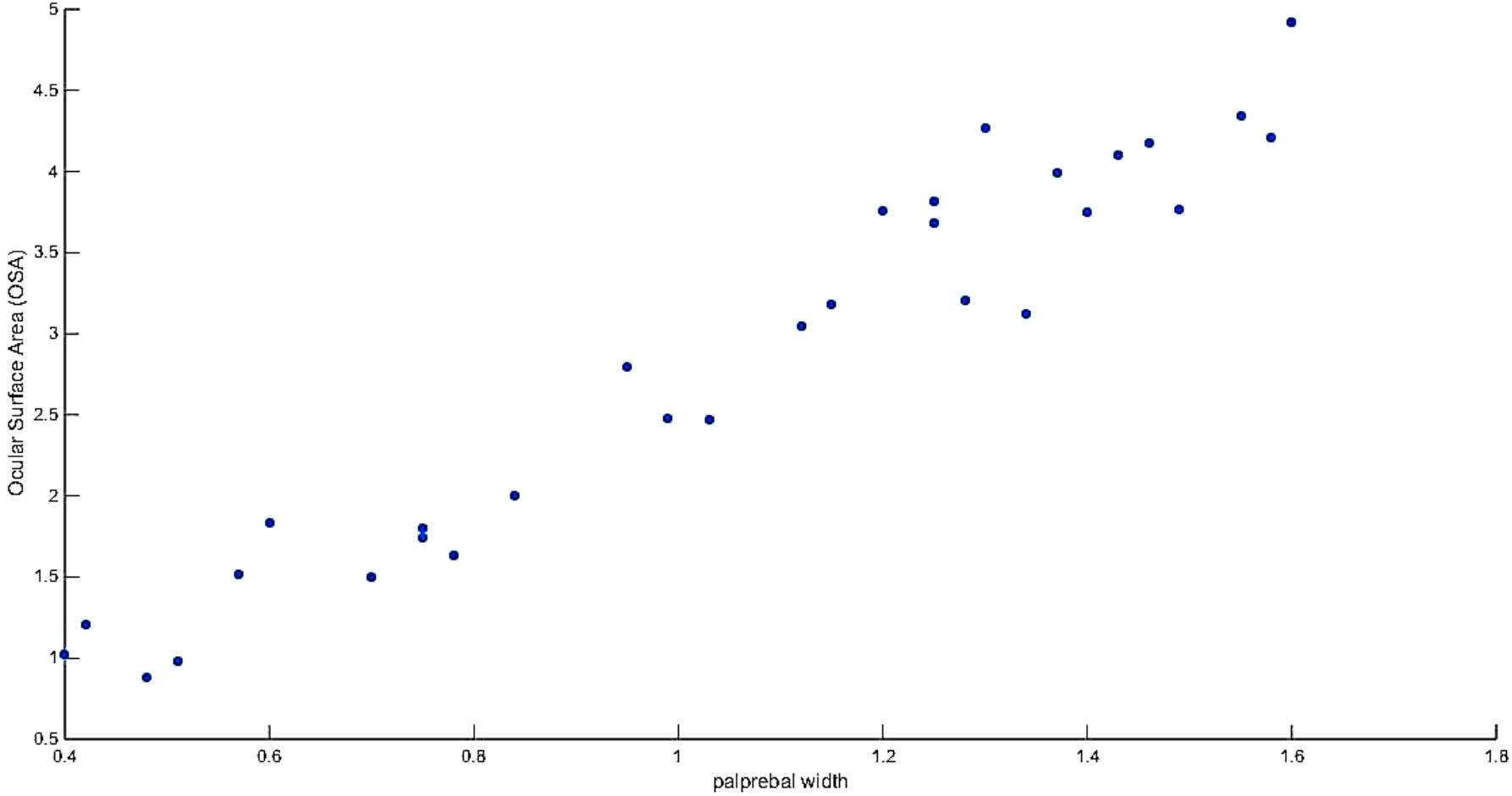
Visual and musculoskeletal problems associated with the use of visual display terminals (VDTs) have become rather common in recent years. Some researchers have focused on vertical gaze direction as a source of eye strain and irritation. This direction is known to be closely related to ocular surface area (OSA), so a method of measuring OSA is needed. The accompanying representative data on y = OSA (cm^2) and x = width of the palprebal fissure (i.e., the horizontal width of the eye opening, in cm) is from the article "Analysis of Ocular Surface Area for Comfortable VDT Workstation Layout" (*Ergonomics,* 1996: 877–884). The order in which observations were obtained was not given, so for convenience they are listed in increasing order of x values.

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
x_i	.40	.42	.48	.51	.57	.60	.70	.75	.75	.78	.84	.95	.99	1.03	1.12
y_i	1.02	1.21	.88	.98	1.52	1.83	1.50	1.80	1.74	1.63	2.00	2.80	2.48	2.47	3.05

i	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
x_i	1.15	1.20	1.25	1.25	1.28	1.30	1.34	1.37	1.40	1.43	1.46	1.49	1.55	1.58	1.60
y_i	3.18	3.76	3.68	3.82	3.21	4.27	3.12	3.99	3.75	4.10	4.18	3.77	4.34	4.21	4.92

Credits: Probability and Statistics for Engineering and the Science by Jan Devore

Scatter Plot Example



Scatter Plot Example: Linear Regression

- Several observations have identical x values yet different y values.
  - *e.g., x_8=x_9 =0 .75, but y_8 =1.80 and y_9 = 1.74). Thus the value of y is not determined solely by x but also by various other factors.*
- There is a strong tendency for y to increase as x increases. That is, larger values of OSA tend to be associated with larger values of fissure width—a positive relationship between the variables.

- It appears that the value of y could be predicted from x by *finding a line that is reasonably close to the points in the plot* (the authors of the cited article superimposed such a line on their plot).

- In other words, there is evidence of a substantial (though not perfect) linear relationship between the two variables.

Simple Linear Regression Model

- For the deterministic model , the actual observed value of y is a linear function of x.
- The appropriate generalization of this to a probabilistic model assumes that the expected value of Y is a linear function of x, but that for fixed x the variable Y differs from its expected value by a random amount.
- In a simple regression problem (a single x and a single y), the form of the model would be:

$$y = b_0 + b_1 * x$$

- In higher dimensions when we have more than one input (x), the line is called a plane or a hyper-plane. The representation therefore is the form of the equation and the specific values used for the coefficients (e.g. b_0 and b_1 in the above example).

Simple Linear Regression Model

- When a coefficient becomes zero, it effectively removes the influence of the input variable on the model and therefore from the prediction made from the model (0 * x = 0).

- The values b_0 and b_1 must be chosen so that they minimize the error. If sum of squared error is taken as a metric to evaluate the model, then goal to obtain a line that best reduces the error.

$$\text{Error} = \sum_{i=1}^{n} (actual\_output - predicted\_output) ** 2$$

Simple Linear Regression Model

- For model with one predictor,

$$b_0 = \bar{y} - b_1\bar{x}$$

and

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Exploring ‘b1’:
  - If b1 > 0, then x(predictor) and y (target) have a positive relationship. That is increase in x will increase y.
  - If b1 < 0, then x(predictor) and y (target) have a negative relationship. That is increase in x will decrease y.

Credits: **Linear Regression — Detailed View by** Saishruthi Swaminathan

Simple Linear Regression Model

- Exploring ‘b0’
  - If the model does not include x=0, then the prediction will become meaningless with only b0.
  - For example, we have a dataset that relates height(x) and weight(y). Taking x=0(that is height as 0), will make equation have only b0 value which is completely meaningless as in real-time height and weight can never be zero. This resulted due to considering the model values beyond its scope.
  - If the model includes value 0, then ‘b0’ will be the average of all predicted values when x=0. But, setting zero for all the predictor variables is often impossible.
  - The value of b0 guarantee that residual have mean zero. If there is no ‘b0’ term, then regression will be forced to pass over the origin. Both the regression co-efficient and prediction will be biased.

Credits: **Linear Regression — Detailed View by** Saishruthi Swaminathan

Slides Adapted
from

Stefano Ermon

Renewable electricity generation in the U.S

	Hydropower	Solar[1]	Wind	Geothermal	Biomass	Total Renewables
2004	6.7%	0.0%	0.4%	0.4%	1.3%	8.8%
2005	6.7%	0.0%	0.4%	0.4%	1.3%	8.8%
2006	7.1%	0.0%	0.7%	0.4%	1.3%	9.5%
2007	5.9%	0.0%	0.8%	0.4%	1.3%	8.5%
2008	6.2%	0.1%	1.3%	0.4%	1.3%	9.3%
2009	6.9%	0.1%	1.9%	0.4%	1.4%	10.6%
2010	6.3%	0.1%	2.3%	0.4%	1.4%	10.4%
2011	7.8%	0.2%	2.9%	0.4%	1.4%	12.6%
2012	6.8%	0.3%	3.4%	0.4%	1.4%	12.4%
2013	6.6%	0.5%	4.1%	0.4%	1.5%	13.1%
2014	6.3%	0.8%	4.4%	0.4%	1.6%	13.5%

Source: Renewable energy data book, NREL

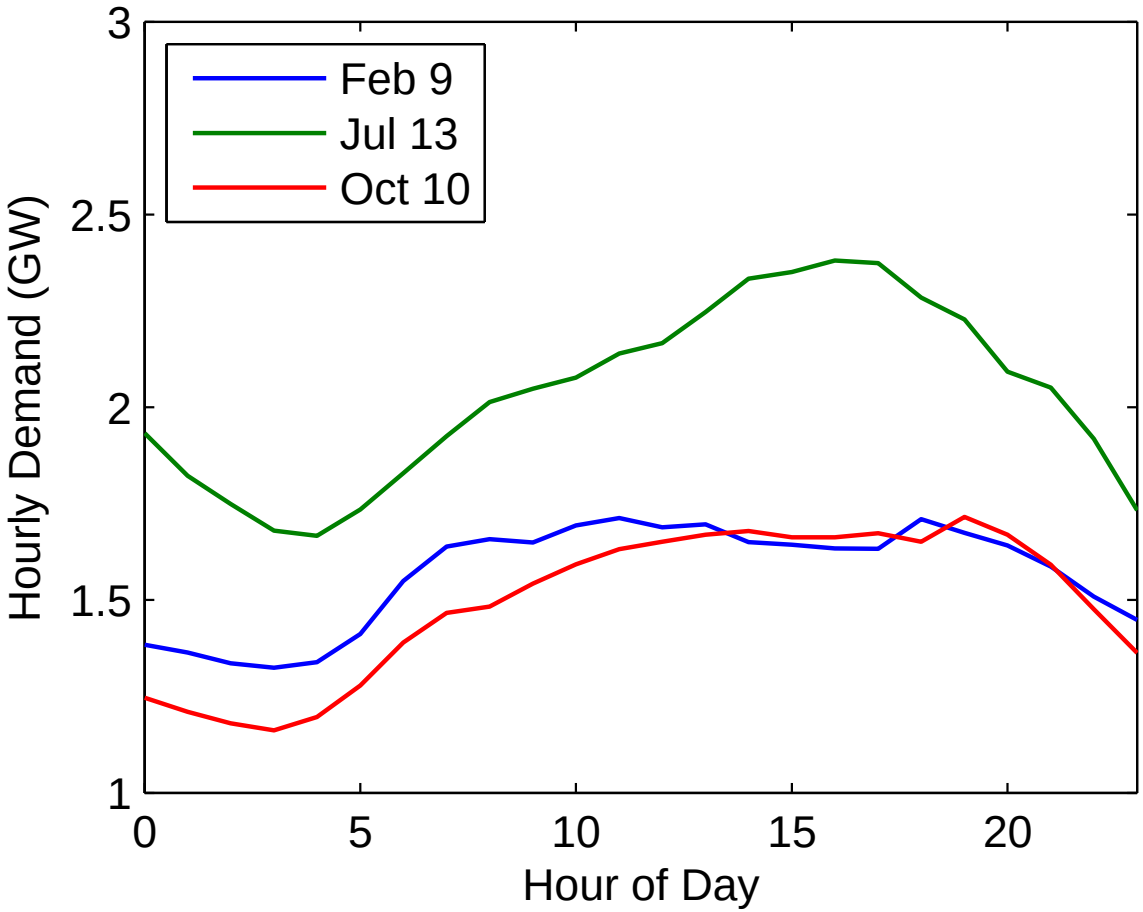
Challenges for the grid

- Wind and solar are intermittent
- We will need traditional power plants when the wind stops
  - Many power plants (e.g., nuclear) cannot be easily turned on/off or quickly ramped up/down
- With more accurate forecasts, wind and solar power become more efficient alternatives
  - A few years ago, Xcel Energy (Colorado) ran ads opposing a proposal that it use 10% of renewable sources
  - Thanks to wind forecasting (ML) algorithms developed at NCAR, they now aim for 30 percent. Accurate forecasting saved the utility $6-$10 million per year

Motivation

- Solar and wind are intermittent
- Can we accurately forecast how much energy will we consume tomorrow?
  - Difficult to estimate from "a priori" models
  - But, we have lots of data from which to build a model

Typical electricity consumption

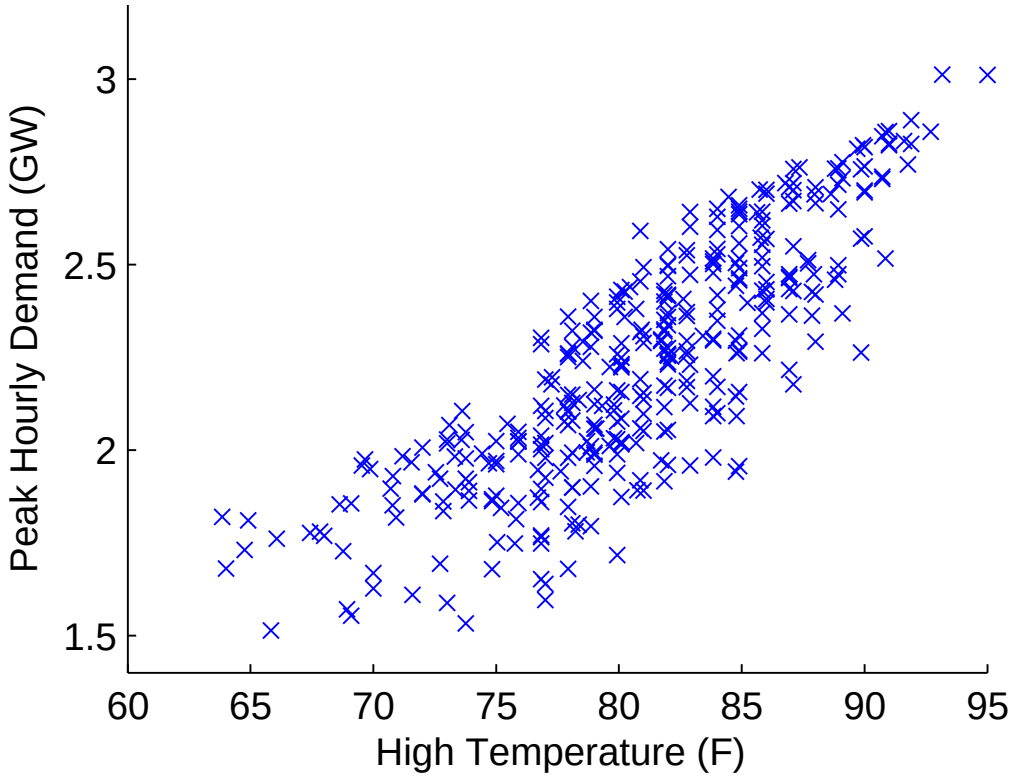


Data: PJM `http://www.pjm.com`

Predict peak demand from high temperature

- What will peak demand be tomorrow?

- If we know something else about tomorrow (like the high temperature), we can use this to *predict* peak demand

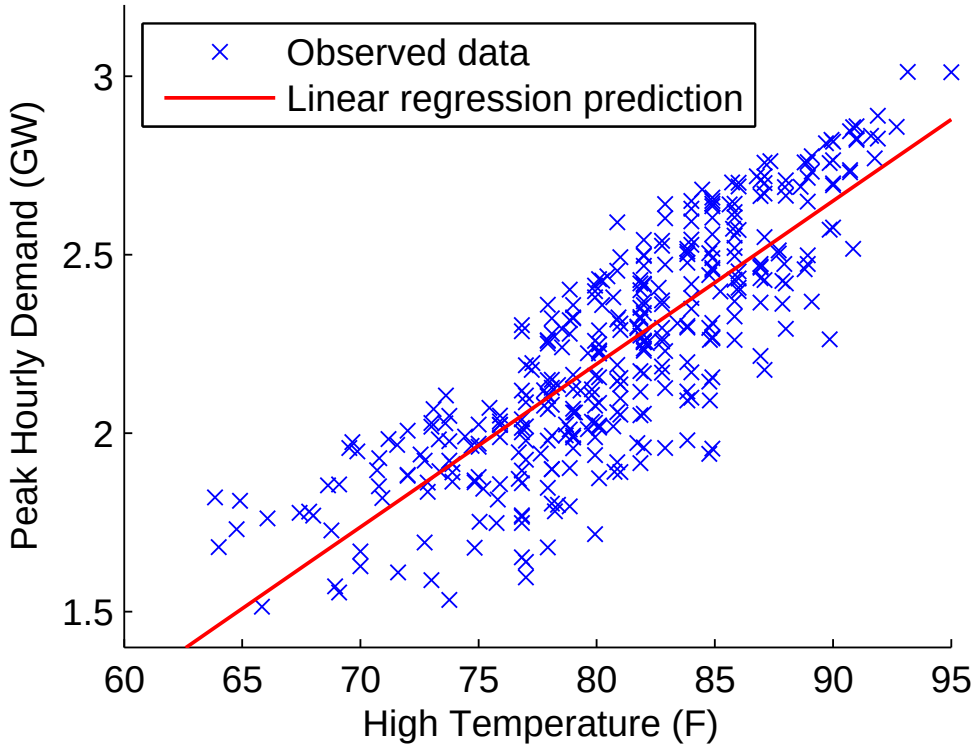


Data: PJM, Weather Underground (summer months, June-August)

A simple model

- A linear model that predicts demand:

$$\text{predicted peak demand} = \theta_1 \cdot (\text{high temperature}) + \theta_2$$



- *Parameters* of model: $\theta_1, \theta_2 \in \mathbb{R}$ $(\theta_1 = 0.046,\ \theta_2 = -1.46)$

A simple model

- We can use a model like this to make predictions
- What will be the peak demand tomorrow?
  - I know from weather report that high temperature will be $80°$F (ignore, for the moment, that this too is a prediction)
- Then predicted peak demand is:

$$\theta_1 \cdot 80 + \theta_2 = 0.046 \cdot 80 - 1.46 = 2.19 \text{ GW}$$

Formal problem setting

- **Input**: $x_i \in \mathbb{R}^n, \quad i = 1, \ldots, m$
  - E.g.: $x_i \in \mathbb{R}^1 = \{\text{high temperature for day } i\}$
- **Output**: $y_i \in \mathbb{R}$ (*regression* task)
  - E.g.: $y_i \in \mathbb{R} = \{\text{peak demand for day } i\}$
- **Model Parameters**: $\theta \in \mathbb{R}^k$
- **Predicted Output**: $\hat{y}_i \in \mathbb{R}$

$$\text{E.g.:} \quad \hat{y}_i = \theta_1 \cdot x_i + \theta_2$$

- For convenience, we define a function that maps inputs to *feature vectors*

$$\phi : \mathbb{R}^n \to \mathbb{R}^k$$

- For example, in our task above, if we define

$$\phi(x_i) = \begin{bmatrix} x_i \\ 1 \end{bmatrix} \quad \text{(here } n = 1,\ k = 2\text{)}$$

then we can write

$$\hat{y}_i = \sum_{j=1}^{k} \theta_j \cdot \phi_j(x_i) \equiv \theta^T \phi(x_i)$$

Loss functions

- Want a model that performs "well" on the data we have

$$\text{I.e., } \hat{y}_i \approx y_i, \quad \forall i$$

- We measure "closeness" of $\hat{y}_i$ and y_i using *loss function*

$$\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$$

- Example: squared loss

$$\ell(\hat{y}_i, y_i) = (\hat{y}_i - y_i)^2$$

Finding model parameters, and optimization

- Want to find model parameters such that minimize sum of costs over all input/output pairs

$$J(\theta) = \sum_{i=1}^{m} \ell(\hat{y}_i, y_i) = \sum_{i=1}^{m} (\theta^T \phi(x_i) - y_i)^2$$

- Write our objective formally as

$$\underset{\theta}{\text{minimize}} \quad J(\theta)$$

simple example of an *optimization problem*; these will dominate our development of algorithms throughout the course

How do we optimize a function

- Search algorithm: Start with an initial guess for θ. Keep changing θ (by a little bit) to reduce $J(\theta)$

- Animation `https://www.youtube.com/watch?v=vWFjqgb-ylQ`

Gradient descent

- Search algorithm: Start with an initial guess for θ. Keep changing θ (by a little bit) to reduce $J(\theta)$

$$J(\theta) = \sum_{i=1}^{m} \ell(\hat{y}_i, y_i) = \sum_{i=1}^{m} (\theta^T \phi(x_i) - y_i)^2$$

- Gradient descent: $\theta_j = \theta_j - \alpha \frac{\partial J(\theta)}{\partial \theta_j}$, for all j

$$\frac{\partial J}{\partial \theta_j} = \frac{\partial \sum_{i=1}^{m} (\theta^T \phi(x_i) - y_i)^2}{\partial \theta_j} = \sum_{i=1}^{m} \frac{\partial (\theta^T \phi(x_i) - y_i)^2}{\partial \theta_j}$$

$$= \sum_{i=1}^{m} 2(\theta^T \phi(x_i) - y_i) \frac{\partial (\theta^T \phi(x_i) - y_i)}{\partial \theta_j}$$

$$= \sum_{i=1}^{m} 2(\theta^T \phi(x_i) - y_i) \phi(x_i)_j$$

Gradient descent

- Repeat until "convergence":

$$\theta_j = \theta_j - \alpha \sum_{i=1}^{m} 2(\theta^T \phi(x_i) - y_i)\phi(x_i)_j, \text{ for all } j$$

- Demo:
  `https://lukaszkujawa.github.io/gradient-descent.html`

- Stochastic gradient descent

- Let's write $J(\theta)$ a little more compactly using matrix notation; define

$$\Phi \in \mathbb{R}^{m \times k} = \begin{bmatrix} — & \phi(x_1)^T & — \\ — & \phi(x_2)^T & — \\ & \vdots & \\ — & \phi(x_m)^T & — \end{bmatrix}, \quad y \in \mathbb{R}^m = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

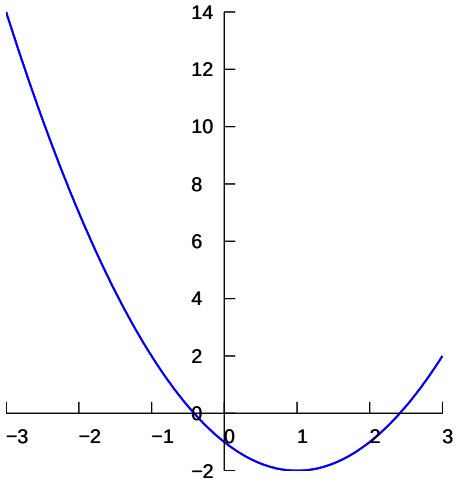
then

$$J(\theta) = \sum_{i=1}^{m} (\theta^T \phi(x_i) - y_i)^2 = \|\Phi\theta - y\|_2^2$$

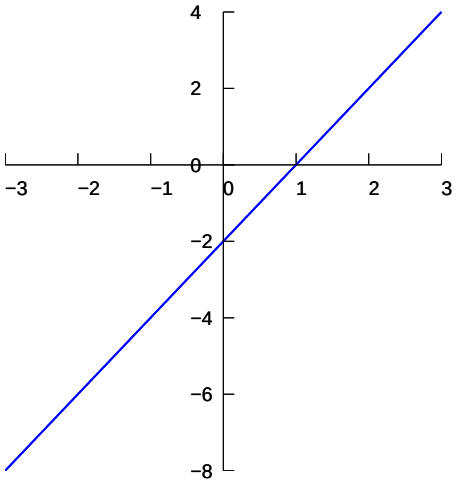
($\|z\|_2$ is ℓ_2 norm of a vector: $\|z\|_2 \equiv \sqrt{\sum_{i=1}^m z_i^2} = \sqrt{z^T z}$)

- Called *least-squares* objective function

- How do we optimize a function? 1-D case ($\theta \in \mathbb{R}$):



$$J(\theta) = \theta^2 - 2\theta - 1$$



$$\frac{dJ}{d\theta} = 2\theta - 2$$

$$\begin{aligned} \theta^\star \text{ minimum} &\implies \left.\frac{dJ}{d\theta}\right|_{\theta^\star} = 0 \\ &\implies 2\theta^\star - 2 = 0 \\ &\implies \theta^\star = 1 \end{aligned}$$

- Multi-variate case: $\theta \in \mathbb{R}^k$, $J : \mathbb{R}^k \to \mathbb{R}$

$$\text{Generalized condition:} \quad \nabla_\theta J(\theta)|_{\theta^\star} = 0$$

- $\nabla_\theta J(\theta)$ denotes *gradient* of J with respect to θ

$$\nabla_\theta J(\theta) \in \mathbb{R}^k \equiv \begin{bmatrix} \frac{\partial J}{\partial \theta_1} \\ \frac{\partial J}{\partial \theta_2} \\ \vdots \\ \frac{\partial J}{\partial \theta_k} \end{bmatrix}$$

- Some important rules and common gradient

$$\nabla_\theta(af(\theta) + bg(\theta)) = a\nabla_\theta f(\theta) + b\nabla_\theta g(\theta), \quad (a, b \in \mathbb{R})$$
$$\nabla_\theta(\theta^T A\theta) = (A + A^T)\theta, \quad (A \in \mathbb{R}^{k\times k})$$
$$\nabla_\theta(b^T\theta) = b, \quad (b \in \mathbb{R}^k)$$

- Optimizing least-squares objective

$$\begin{aligned} J(\theta) &= \|\Phi\theta - y\|_2^2 \\ &= (\Phi\theta - y)^T(\Phi\theta - y) \\ &= \theta^T\Phi^T\Phi\theta - 2y^T\Phi\theta + y^Ty \end{aligned}$$

- Using the previous gradient rules

$$\begin{aligned} \nabla_\theta J(\theta) &= \nabla_\theta(\theta^T\Phi^T\Phi\theta - 2y^T\Phi\theta + y^Ty) \\ &= \nabla_\theta(\theta^T\Phi^T\Phi\theta) - 2\nabla_\theta(y^T\Phi\theta) + \nabla_\theta(y^Ty) \\ &= 2\Phi^T\Phi\theta - 2\Phi^Ty \end{aligned}$$

- Setting gradient equal to zero

$$2\Phi^T\Phi\theta^\star - 2\Phi^Ty = 0 \iff \theta^\star = (\Phi^T\Phi)^{-1}\Phi^Ty$$

known as the *normal equations*

- Let's see how this looks in MATLAB code

```
X = load('high_temperature.txt');
y = load('peak_demand.txt');
n = size(X,2);
m = size(X,1);
Phi = [X ones(m,1)];
theta = inv(Phi' * Phi) * Phi' * y;

theta =
    0.0466
   -1.4600
```

- The normal equations are so common that MATLAB has a special operation for them

```
% same as inv(Phi' * Phi) * Phi' * y
theta = Phi \ y;
```

Higher-dimensional inputs

- Input: $x \in \mathbb{R}^2 = \begin{bmatrix} \text{temperature} \\ \text{hour of day} \end{bmatrix}$
- Output: $y \in \mathbb{R} =$ demand

- Features: $\phi(x) \in \mathbb{R}^3 = \begin{bmatrix} \text{temperature} \\ \text{hour of day} \\ 1 \end{bmatrix}$

- Same matrices as before

$$\Phi \in \mathbb{R}^{m \times k} = \begin{bmatrix} — & \phi(x_1)^T & — \\ & \vdots & \\ — & \phi(x_m)^T & — \end{bmatrix}, \quad y \in \mathbb{R}^m = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}$$

- Same solution as before

$$\theta \in \mathbb{R}^3 = (\Phi^T \Phi)^{-1} \Phi^T y$$