Supervised Learning – Regression

COMP9417 Machine Learning and Data Mining

Last revision: 7 Mar 2018

Acknowledgements

Material derived from slides for the book
"Elements of Statistical Learning (2nd Ed.)" by T. Hastie,
R. Tibshirani & J. Friedman. Springer (2009)
http://statweb.stanford.edu/~tibs/ElemStatLearn/

Material derived from slides for the book
"Machine Learning: A Probabilistic Perspective" by P. Murphy
MIT Press (2012)
http://www.cs.ubc.ca/~murphyk/MLbook

Material derived from slides for the book
"Machine Learning" by P. Flach
Cambridge University Press (2012)
http://cs.bris.ac.uk/~flach/mlbook

Material derived from slides for the book
"Bayesian Reasoning and Machine Learning" by D. Barber
Cambridge University Press (2012)
http://www.cs.ucl.ac.uk/staff/d.barber/brml

Material derived from slides for the book
"Machine Learning" by T. Mitchell
McGraw-Hill (1997)
http://www-2.cs.cmu.edu/~tom/mlbook.html

Material derived from slides for the course
"Machine Learning" by A. Srinivasan
BITS Pilani, Goa, India (2016)

Aims

This lecture will introduce you to machine learning approaches to the problem of numerical prediction. Following it you should be able to reproduce theoretical results, outline algorithmic techniques and describe practical applications for the topics:

- the supervised learning task of numeric prediction
- how linear regression solves the problem of numeric prediction
- fitting linear regression by least squares error criterion
- non-linear regression via linear-in-the-parameters models
- parameter estimation for regression
- local (nearest-neighbour) regression

Note: slides with titles marked * are for background only.

Introduction

In the introductory lecture we saw supervised learning methods mostly for *classification*, where the task is prediction of a *discrete* value for data instances . . .

. . . however, we often find tasks where the most natural representation is that of *prediction of numeric values*

Introduction

Task: learn a model to predict CPU performance from a datset of example of 209 different computer configurations.

	Cycle time (ns)	Main memory (Kb)		Cache (Kb)	Channels		Performance
	MYCT	MMIN	MMAX	CACH	CHMIN	CHMAX	PRP
1	125	256	6000	256	16	128	198
2	29	8000	32000	32	8	32	269
...							
208	480	512	8000	32	0	0	67
209	480	1000	4000	0	0	0	45

Introduction

Result: a linear regression equation fitted to the CPU dataset.


```
PRP =
  - 56.1
    + 0.049 MYCT
    + 0.015 MMIN
    + 0.006 MMAX
    + 0.630 CACH
    - 0.270 CHMIN
    + 1.46 CHMAX
```

Introduction

For the class of *symbolic* representations, machine learning is viewed as:

searching a space of **hypotheses** . . .

represented in a formal hypothesis language (trees, rules, graphs . . .).

Introduction

For the class of *numeric* representations, machine learning is viewed as:

“searching” a space of **functions** . . .

represented as mathematical models (linear equations, neural nets, . . .).

Note: in both settings, the models may be probabilistic . . .

Introduction

Methods to predict a numeric output from statistics and machine learning:

- linear regression (statistics) determining the "line of best fit" using the least squares criterion
- linear models (machine learning) learning a predictive model from data under the assumption of a linear relationship between predictor and target variables

Very widely-used, many applications

Ideas that are generalised in Artificial Neural Networks

Introduction

- non-linear regression by adding non-linear basis functions
- multi-layer neural networks (machine learning) learning non-linear predictors via hidden nodes between input and output

- regression trees (statistics / machine learning) tree where each leaf predicts a numeric quantity
- local (nearest-neighbour) regression

Learning Linear Regression Models

Regression

We will look at the simplest model for numerical prediction:

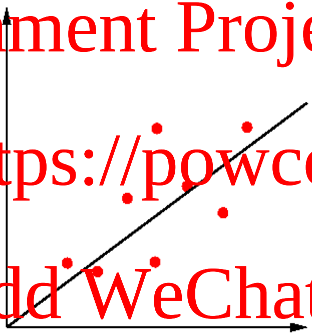
a *regression equation*

Outcome will be a linear sum of feature values with appropriate weights.

Note: the term *regression* is overloaded – it can refer to:

- the process of determining the weights for the regression equation, or
- the regression equation itself.

Linear Regression



inputs	outputs
$x1 = 1$	$y1 = 1$
$x2 = 3$	$y2 = 2.2$
$x3 = 2$	$y3 = 2$
$x4 = 1.5$	$y4 = 1.9$
$x5 = 4$	$y5 = 3.1$

Assumes: expected value of the output given an input, $E[y|x]$, is linear.
Simplest case: $\text{Out}(x) = bx$ for some unknown b.
Learning problem: given the data, estimate b (i.e., $\hat{b}$).

Linear Models

- Numeric attributes and numeric prediction, i.e., regression
- Linear models, i.e. outcome is *linear* combination of attributes

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$$

- Weights are calculated from the training data
- **Predicted** value for first training instance $\mathbf{x}^{(1)}$ is:

$$b_0 x_0^{(1)} + b_1 x_1^{(1)} + b_2 x_2^{(1)} + \ldots + b_n x_n^{(1)} = \sum_{i=0}^{n} b_i x_i^{(1)}$$

Minimizing Squared Error

Difference between *predicted* and *actual* values is the error !

$n+1$ coefficients are chosen so that sum of squared error on all instances in training data is minimized

Squared error:

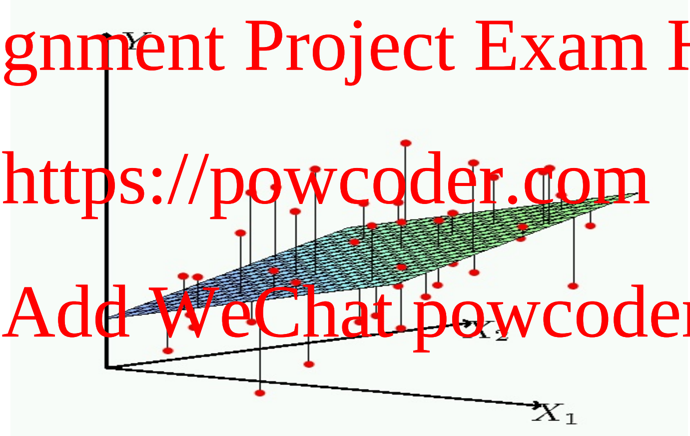
$$\sum_{j=1}^{m}\left(y^{(j)}-\sum_{i=0}^{n} b_i x_i^{(j)}\right)^2$$

Coefficients can be derived using standard matrix operations

Can be done if there are more instances than attributes (roughly speaking).

Known as "Ordinary Least Squares" (OLS) regression – minimizing the sum of squared distances of data points to the estimated regression line.

Multiple Regression



Example: linear least squares fitting with 2 input variables.

Step back: Statistical Techniques for Data Analysis

Probability vs Statistics: The Difference

- **Probability** versus **Statistics**
- Probability: reasons from populations to samples
  - This is deductive reasoning, and is usually *sound* (in the logical sense of the word)

- Statistics: reasons from samples to populations
  - This is inductive reasoning, and is usually *unsound* (in the logical sense of the word)

Statistical Analyses

- Statistical analyses usually involve one of 3 things:
  1. The study of populations;
  2. The study of variation; and
  3. Techniques for data abstraction and data reduction

- Statistical analysis is more than statistical computation:
  1. What is the question to be answered?
  2. Can it be quantitative (i.e., can we make measurements about it)?
  3. How do we collect data?
  4. What can the data tell us?

Sampling

Where do the Data come from? (Sampling)

- For groups (populations) that are fairly homogeneous, we do not need to collect a lot of data. (We do not need to sip a cup of tea several times to decide that it is too hot.)

- For populations which have irregularities, we will need to either take measurements of the entire group, or find some way of get a good idea of the population without having to do so

- *Sampling* is a way to draw conclusions about the population without having to measure all of the population. The conclusions need not be completely accurate
- All this is possible if the sample closely resembles the population about which we are trying to draw some conclusions

What We Want From a Sampling Method

- No systematic bias, or at least no bias that we cannot account for in our calculations

- The chance of obtaining an unrepresentative sample can be calculated. (So, if this chance is high, we can choose not to draw any conclusions.)
- The chance of obtaining an unrepresentative sample decreases with the size of the sample

Simple Random Sampling

- Each element of the population is associated with a number
- Shuffle all the numbers and put them into into a hat
- Draw a sample of n numbers from the hat and get the corresponding elements of the population

Usually, there are no hats :) and we will be using a computer to generate n numbers that are approximately random.

In addition, the computer will use a mathematical relationship between elements of the population and the set of numbers. Inverting this relationship using the n random numbers will then give the elements of the population.

Probability Sampling

- In effect, numbers drawn using simple random sampling (in a single stage or more) use a uniform probability distribution over the numbers. That is, the probability of getting any number from $1 \ldots n$ from the hat is $1/n$.
- A more general form of this is to use any kind of probability distribution over $1 \ldots n$. For example, a distribution could make larger numbers are more likely than smaller numbers. This is a skewed distribution
- For example, take a 2-stage sampling procedure in which households are grouped according to size, and the probability of selecting larger households is higher. A household is selected and then an individual is selected from that household. This gives a greater chance of selecting individuals from larger households
- Once again, it is relatively straightforward to do this form of probability-based sampling using a computer

Estimation

Estimation from a Sample

- Estimating some aspect of the population using a sample is a common task. Along with the estimate, we also want to have some idea of the accuracy of the estimate (usually expressed in terms of *confidence limits*)

- Some measures calculated from the sample are very good estimates of corresponding population values. For example, the sample mean m is a very good estimate of the population mean μ. But this is not always the case. For example, the range of a sample usually under-estimates the range of the population

- We will have to clarify what is meant by a "good estimate". One meaning is that an estimator is correct on average. For example, on average, the mean of a sample is a good estimator of the mean of the population

Estimation from a Sample

- For example, when a number of samples are drawn and the mean of each is found, then average of these means is equal to the population mean

- Such an estimator is said to be *statistically unbiased*

Sample Estimates of the Mean and the Spread I

Mean. This is calculated as follows.

- Find the total T of N observations. Estimate the (arithmetic) mean by $m = T/N$.
- This works very well when the data follow a symmetric bell-shaped frequency distribution (of the kind modelled by "normal" distribution)
- A simple mathematical expression of this is $m = \frac{1}{N}\sum_i x_i$, where the observations are $x_1, x_2 \ldots x_n$
- If we can group the data so that the observation x_1 occurs f_1 times, x_2 occurs f_2 times and so on, then the mean is calculated even easier as $m = \frac{1}{N}\sum_i x_i f_i$

Sample Estimates of the Mean and the Spread II

- If, instead of frequencies, you had relative frequencies (i.e. instead of f_i you had $p_i = f_i/N$), then the mean is simply the observations weighted by relative frequency. That is, $m = \sum_i x_i p_i$

- We want to connect this up to computing the mean value of observations modelled by some theoretical probability distribution function. That is, we want to a similar counting method for calculating the mean of random variables modelled using some known distribution

Sample Estimates of the Mean and the Spread III

- Correctly, this is the mean value of the *values of the random variable function*. But this is a bit cumbersome, so we will just say the "mean value of the r.v." For discrete r.v.'s this is:

$$E(X) = \sum_i x_i p(X = x_i)$$

Variance. This is calculated as follows:

- Calculate the total T and the sum of squares of N observations. The estimate of the standard deviation is $s = \sqrt{\frac{1}{N-1}\sum_i (x_i - m)^2}$
- Again, this is a very good estimate when the data are modelled by a normal distribution

Sample Estimates of the Mean and the Spread IV

- For grouped data, this is modified to
  $s = \sqrt{\frac{1}{N-1}\sum_i (x_i - m)^2 f_i}$
- Again, we have a similar formula in terms of expected values, for the scatter (spread) of values of a r.v. X around a mean value $E(X)$:

$$\begin{aligned} Var(X) &= E(X - E(X))^2 \\ &= E(X^2) - [E(X)]^2 \end{aligned}$$

- You can remember this as "the mean of the squares minus the square of the mean"

Bias-Variance Decomposition

The Bias-Variance Tradeoff

- When comparing unbiased estimators, we would like to select the one with minimum variance
- In general, we would be comparing estimators that have some bias and some variance
- We can combine the bias and variance of an estimator by obtaining the *mean square error* of the estimator, or MSE. This is the average value of squared deviations of an estimated value V from the true value of the parameter θ. That is:

$$\text{MSE} = \text{Avg. value of } (V - \theta)^2$$

- Now, it can be shown that:

$$\text{MSE} = (\text{variance}) + (\text{bias})^2$$

- If, as sample size increases, the bias and the variance of an estimator approaches 0, then the estimator is said to be *consistent*.

The Bias-Variance Tradeoff

- Since

$$\text{MSE} = (\text{variance}) + (\text{bias})^2$$

the lowest possible value of MSE is 0
- In general, we may not be able to get to the ideal MSE of 0. Sampling theory tells us the minimum value of the variance of an estimator. This value is known as the *Cramer-Rao* bound. So, given an estimator with bias b, we can calculate the minimum value of the variance of the estimator using the CR bound (say, v_{min}). Then:

$$\text{MSE} \geq v_{min} + b^2$$

The value of v_{min} depends on whether the estimator is biased or unbiased (that is $b = 0$ or $b \neq 0$)
- It is not the case that v_{min} for an unbiased ($b = 0$) estimator is less than v_{min} for a biased estimator. So, the MSE of a biased estimator can end up being lower than the MSE of an unbiased estimator.

Decomposition of MSE

$$\mathrm{MSE} = (\text{variance}) + (\text{bias})^2$$

Imagine testing the prediction of our estimator $\hat{y}$ on many samples of the same size drawn at random from the same distribution. We compute error based on the squared difference between predicted and actual values. Then the MSE can be decomposed like this:

$$\begin{aligned}\mathrm{MSE} &= E[\hat{y} - f(x)]^2 \\ &= E[\hat{y} - E(\hat{y})]^2 + [E(\hat{y}) - f(x)]^2\end{aligned}$$

Note that the first term in the error decomposition (variance) does not refer to the actual value at all, although the second term (bias) does.

Correlation

Correlation I

- The *correlation coefficient* is a number between -1 and +1 that indicates whether a pair of variables x and y are associated or not, and whether the scatter in the association is high or low
  - High values of x are associated with high values of y *and* low values of x are associated with low values of y, and scatter is low
  - A value near 0 indicates that there is no particular association and that there is a large scatter associated with the values
  - A value close to -1 suggests an inverse association between x and y
- Only appropriate when x and y are roughly linearly associated (doesn't work well when the association is curved)
- The formula for computing correlation between x and y is:

$$r = \frac{\mathrm{cov}(x, y)}{\sqrt{\mathrm{var}(x)}\sqrt{\mathrm{var}(y)}}$$

This is sometimes also called *Pearson's correlation coefficient*

Correlation II

- The terms in the denominator are simply the standard deviations of x and y. But the numerator is different. This is calculated as the average of the product of deviations from the mean:

$$\mathrm{cov}(x, y) = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{n - 1}$$

- What does "covariance" mean ? Consider

1. Case 1: $x_i > \overline{x}$, $y_i > \overline{y}$
2. Case 2: $x_i < \overline{x}$, $y_i < \overline{y}$
3. Case 3: $x_i < \overline{x}$, $y_i > \overline{y}$
4. Case 4: $x_i > \overline{x}$, $y_i < \overline{y}$

In the first two cases, x_i and y_i vary together, both being high or low relative to their means. In the other two cases, they vary in different directions

Correlation III

- If the positive products dominate in the calculation of $\text{cov}(x, y)$, then the value of r will be positive. If the negative products dominate, then r will be negative. If 0 products dominate, then r will be close to 0.
- You should be able to show that:

$$r = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2}\sqrt{\sum (y_i - \overline{y})^2}}$$

- Computers generally use a short-cut formula:

$$r = \frac{\sum_i x_i y_i - n\overline{x}\overline{y}}{n - 1}$$

- The same kinds of calculations can be done if the data were not actual values but ranks instead (i.e. ranks for the x's and the y's).
  - This is called *Spearman's rank correlation*, but we won't do these calculations here.

What Happens If You Sample? I

- Suppose you have a sample of $\langle x, y \rangle$ pairs and you calculate $r = 0.3$. Is this really the case?
- Sampling theory tells us something. If: (a) the relative frequencies observed are well modelled by a special kind of mathematical function (a "Normal" or Gaussian distribution); (b) the true correlation is 0; and (c) the number of samples is large

- Then:
  - The sampling distribution of the correlation coefficient (that is, how r varies from sample to sample) is also approximately distributed according to the Normal distribution with mean 0 and standard error (s.e.) of approximately $1/\sqrt{n}$

- We can use this to calculate the (approximate) probability of obtaining the sample if the assumptions were true

What Happens If You Sample? II

- Suppose we calculate $r = 0.3$ from the sample, and that the s.e. is 0.1, say. Then if the sample came from a population with true correlation 0, this would be quite unusual (less than 1% chance)

- We would say instead that the sample was probably from a population with correlation 0.3, with a 95% confidence interval of $\pm 2 \times 0.1$

What Does Correlation Mean? I

- r is a quick way of checking whether there is some linear association between x and y

- The sign of the value tells you the direction of the association
- All that the numerical value tells you is about the scatter in the data
- The correlation coefficient does not model any relationship. That is, given a particular x you cannot use the r value to calculate a y value

    - It is possible for two datasets to have the same correlation, but different relationships
    - It is possible for two datasets to have different correlations but the same relationship

- MORAL: Do not use correlations to compare datasets. All you can derive is whether there is a positive or negative relationship between x and y
- ANOTHER MORAL: Do not use correlation to imply x causes y or the other way around

Regression

Regression

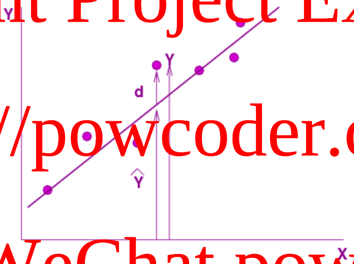
- Given a set of data points x_i, y_i, what is the relationship between them? (We can generalise this to the "multivariate" case later)
- One kind of question is to ask: are these linearly related in some manner? That is, can we draw a straight line that describes reasonably well the relationship between X and Y

- Remember, the correlation coefficient can tell us if there is a case for such a relationship

- In real life, even if such a relationship held, it will be unreasonable to expect all pairs x_i, y_i to lie precisely on a straight line. Instead, we can probably draw some reasonably well-fitting line. But which one?

Univariate linear regression

Linear Relationship Between 2 Variables I



- GOAL: fit a line whose equation is of the form $\hat{Y} = a + bX$
- HOW: minimise $\sum_i d_i^2 = \sum_i (Y_i - \hat{Y}_i)^2$ (the "least squares estimator")

Linear Relationship Between 2 Variables II

- The calculation for b is given by:

$$b = \frac{\text{cov}(x,y)}{\text{var}(x)}$$

where $\text{cov}(x,y)$ is the covariance of x and y, given by: $\sum_i (x_i - \bar{x})(y_i - \bar{y})$ as before

- This can be simplified to:

$$b = \sum(xy) / \sum x^2$$

where $x = (X_i - \overline{X})$ and $y = (Y_i - \overline{Y})$

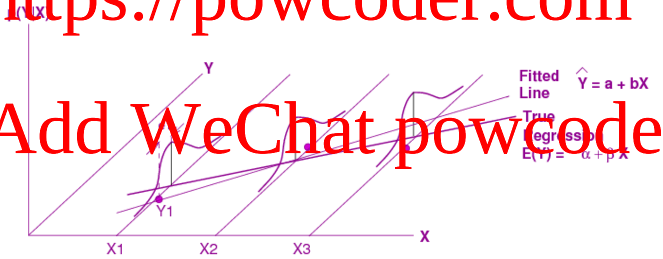
- $a = \overline{Y} - b\overline{X}$

Meaning of the Coefficients a and b

- b: change in Y that accompanies a unit change in X
- If the values of X were assigned at random, then b estimates the unit change in Y *caused* by a unit change in X
- If the values of X were not assigned at random (for examples, they were data somebody observed), then the change in Y will include the change in X and any other confounding variables that may have changed as a result of changing X by 1 unit. So, you cannot say for example, that a change of X by 1 unit causes b units of change in Y
- $b = 0$ means there is no linear relationship between X and Y, and then best we can do is simply say is $\hat{Y} = a = \overline{Y}$. Estimating the sample mean is therefore a special case of the MSE criterion

The Regression Model I

- The least-square estimator fits a line using sample data
- To draw inferences about the population requires us to have a (statistical) model about what this line means
- What is being assumed is actually this:



The Regression Model II

- That is: Obtain Y values for many instances of X_1. This will result in a distribution of Y values $P(Y|X_1)$; and so on for $P(Y|X_2), P(Y|X_3), etc.$. The regression model makes the following assumptions:
  - All the Y distributions are the same, and have the same spread
  - For each $P(Y|X_i)$ distribution, the true mean value μ_i lies on a straight line (this is the "true regression line")
  - The Y_i are independent
- In standard terminology, the Y_i are identically distributed independent (i.i.d.) random variables with mean $\mu_i = \alpha + \beta X_i$ and variance σ^2
- Or: $Y_i = \alpha + \beta X_i + e_i$ where the e_i are independent errors with mean 0 and variance σ^2

How Good is the Least-Squares Estimator I

- The line fitted using the least-squares criterion is a sample-based estimate of the true regression line
- To know how good this estimate is, we are really asking questions about the bias and variance of the estimates of a and b

- It can be shown that under some assumptions, the least-square estimates of a and b will be unbiased and that they will have the lowest variance
- The proof of this is called the *Gauss-Markov theorem*. The Gauss-Markov theorem makes the following assumptions

1. The expected (average) values of residuals is 0 ($E(e_i) = 0$)
2. The spread of residuals is constant for all X_i ($Var(e_i) = \sigma^2$)
3. There is no relationship amongst the residuals ($cov(e_i, e_j) = 0$)
4. There is no relationship between the residuals and the X_i ($cov(X_i, e_i) = 0$)

How Good is the Least-Squares Estimator II

- If these assumptions hold, then the Gauss-Markov theorem shows that $E(a) = \alpha$, $E(b) = \beta$, and that the variance in these estimates will have the lowest variance

- There is a special case of the assumptions that arises when the residuals are assumed to be distributed according to the Normal distribution, with mean 0
  - In this case, minimising least-squares is equivalent to maximising the probability of the Y_i, given the X_i (that is, least-squares is equivalent to *maximum likelihood estimation*)
  - More on this in a later lecture

Univariate linear regression

Example:

Suppose we want to investigate the relationship between people's height and weight. We collect n height and weight measurements $(h_i, w_i), 1 \leq i \leq n$.

Univariate linear regression assumes a linear equation $w = a + bh$, with parameters a and b chosen such that the sum of squared residuals $\sum_{i=1}^{n}(w_i - (a + bh_i))^2$ is minimised.

Univariate linear regression

In order to find the parameters we take partial derivatives, set the partial derivatives to 0 and solve for a and b:

$$\frac{\partial}{\partial a}\sum_{i=1}^{n}(w_i-(a+bh_i))^2=-2\sum_{i=1}^{n}(w_i-(a+bh_i))=0$$

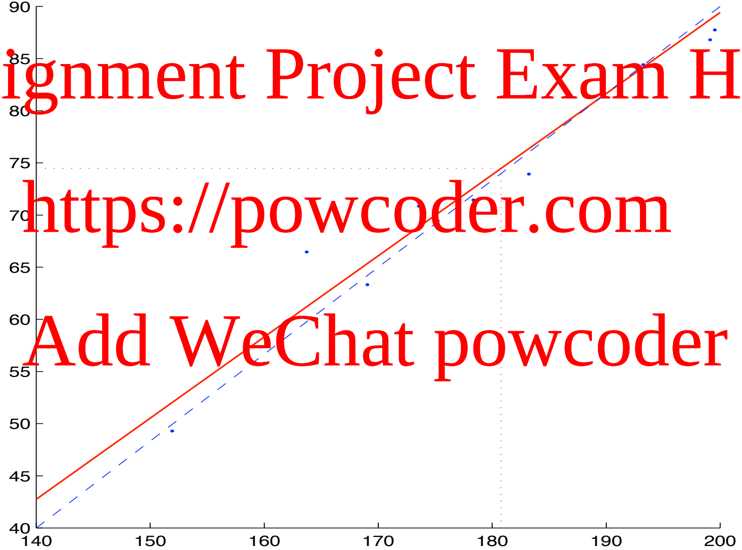
$$\Rightarrow \hat{a}=\overline{w}-\hat{b}\overline{h}$$

$$\frac{\partial}{\partial b}\sum_{i=1}^{n}(w_i-(a+bh_i))^2=-2\sum_{i=1}^{n}(w_i-(a+bh_i))h_i=0$$

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

So the solution found by linear regression is $w=\hat{a}+\hat{b}h=\overline{w}+\hat{b}(h-\overline{h})$.

Univariate linear regression



Univariate linear regression

Shown on previous slide:

The red solid line indicates the result of applying linear regression to 10 measurements of body weight (on the y-axis, in kilograms) against body height (on the x-axis, in centimetres). The orange dotted lines indicate the average height $\overline{h} = 181$ and the average weight $\overline{w} = 74.5$; the regression coefficient $\hat{b} = 0.78$. The measurements were simulated by adding normally distributed noise with mean 0 and variance 5 to the true model indicated by the blue dashed line ($b = 0.83$).

Linear regression: intuitions

For a feature x and a target variable y, the regression coefficient is the covariance between x and y in proportion to the variance of x:

$$\hat{b} = \frac{\sigma_{xy}}{\sigma_{xx}}$$

(Here we use σ_{xx} as an alternative notation for σ_x^2).

This can be understood by noting that the covariance is measured in units of x times units of y (e.g., metres times kilograms above) and the variance in units of x squared (e.g., metres squared), so their quotient is measured in units of y per unit of x (e.g., kilograms per metre).

Linear regression: intuitions

The *intercept* $\hat{a}$ is such that the regression line goes through $(\overline{x}, \overline{y})$.

Adding a constant to all x-values (a translation) will affect only the intercept but not the regression coefficient (since it is defined in terms of deviations from the mean, which are unaffected by a translation).

So we could *zero-centre* the x-values by subtracting $\overline{x}$, in which case the intercept is equal to $\overline{y}$.

We could even subtract $\overline{y}$ from all y-values to achieve a zero intercept, without changing the problem in an essential way.

Linear regression: intuitions

Suppose we replace x_i with $x'_i = x_i / \sigma_{xx}$ and likewise $\overline{x}$ with $\overline{x'} = \overline{x} / \sigma_{xx}$, then we have that $\hat{b} = \frac{1}{n}\sum_{i=1}^{n}(x'_i - \overline{x'})(y_i - \overline{y}) = \sigma_{x'y}$.

In other words, if we *normalise* x by dividing all its values by x's variance, we can take the covariance between the normalised feature and the target variable as regression coefficient.

This demonstrates that univariate linear regression can be understood as consisting of two steps:

1. normalisation of the feature by dividing its values by the feature's variance;
2. calculating the covariance of the target variable and the normalised feature.

Linear regression: intuitions

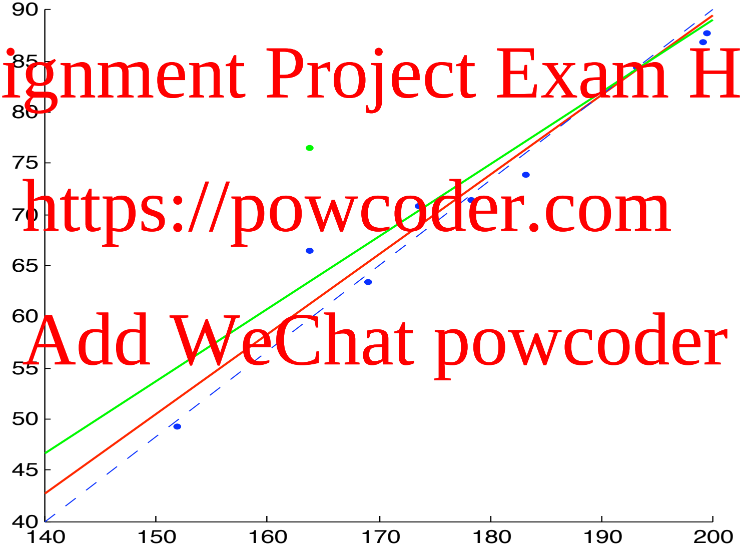
Another important point to note is that the sum of the residuals of the least-squares solution is zero:

$$\sum_{i=1}^{n}(y_i - (\hat{a} + \hat{b}x_i)) = n(\overline{y} - \hat{a} - \hat{b}\overline{x}) = 0$$

The result follows because $\hat{a} = \overline{y} - \hat{b}\overline{x}$, as derived above.

While this property is intuitively appealing, it is worth keeping in mind that it also makes linear regression susceptible to *outliers*: points that are far removed from the regression line, often because of measurement errors.

The effect of outliers



The effect of outliers

Shown on previous slide:

Suppose that, as the result of a transcription error, one of the weight values from the previous example of univariate regression is increased by 10 kg. The diagram shows that this has a considerable effect on the least-squares regression line.

Specifically, we see that one of the blue points got moved up 10 units to the green point, changing the red regression line to the green line.

Least-Squares as Cost Minimization I

- Finding the least-squares solution is in effect finding the value of a and b that minimizes $\sum_i d_i^2 = \sum_i (Y_i - \hat{Y}_i)^2$, where $\hat{Y}_i = a + bX_i$
- This minimum value was obtained analytically by the usual process of differentiating and equating to 0,
- A numerical alternative to the analytical approach is to take (small) steps that decreases the value of the function to be minimised, and stopping when we reach a minimum
- Recall that at a point the gradient vector points in the direction of greatest increase of a function. So, the opposite direction to the gradient vector gives the direction of greatest decrease
  - $b_{i+1} = b_i$ - $\eta \times g_b$
  - $a_{i+1} = a_i$ - $\eta \times g_a$
  - Stop when $b_{i+1} \approx b_i$ and $a_{i+1} \approx a_i$
- More on this in a later lecture

Multivariate linear regression

Many variables

- Often, we are interesting in modelling the relationship of Y to several other variables

- In observational studies, the value of Y may be affected by the values of several variables. For example, carcinogenicity may be gender-specific. A regression model that ignores gender may find that carcinogenicity to be related to some surrogate variable (height, for example)

- Including more variables can give a narrower confidence interval on the prediction being made

Multivariate linear model

- The Y_i are identically distributed independent variables with mean $\mu = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$ and variance σ^2
- Or: $Y_i = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + e_i$ where the e_i are independent errors with mean 0 and variance σ^2
- As before, this linear model is estimated from a sample by the equation $\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_n X_n$
- With many variables, the regression equation and expressions for the b_i are expressed better using a matrix representation for sets of equations.

Multivariate linear regression *

First, we need the covariances between every feature and the target variable:

$$(\mathbf{X}^{\mathrm{T}}\mathbf{y})_j = \sum_{i=1}^{n} x_{ij} y_i = \sum_{i=1}^{n} (x_{ij} - \mu_j)(y_i - \overline{y}) + n\mu_j \; \overline{y} = n(\sigma_{jy} + \mu_j \; \overline{y})$$

Assuming for the moment that every feature is zero-centred, we have $\mu_j = 0$ and thus $\mathbf{X}^{\mathrm{T}}\mathbf{y}$ is an n-vector holding all the required covariances (times n).

We can normalise the features by means of a d-by-d *scaling matrix*: a *diagonal matrix* with diagonal entries $1/n\sigma_{jj}$. If $\mathbf{S}$ is a diagonal matrix with diagonal entries $n\sigma_{jj}$, we can get the required scaling matrix by simply inverting $\mathbf{S}$.

So our first stab at a solution for the *multivariate regression* problem is

$$\hat{\mathbf{w}} \;\; = \;\; \mathbf{S}^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

Multivariate linear regression *

The general case requires a more elaborate matrix instead of $\mathbf{S}$:

$$\hat{\mathbf{w}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \tag{1}$$

Let us try to understand the term $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ a bit better.

- Assuming the features are uncorrelated, the *covariance matrix* $\mathbf{\Sigma}$ is diagonal with entries σ_{jj}.
- Assuming the features are zero-centred, $\mathbf{X}^{\mathrm{T}}\mathbf{X} = n\mathbf{\Sigma}$ is also diagonal with entries $n\sigma_{jj}$.
- In other words, assuming zero-centred and uncorrelated features, $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ reduces to our scaling matrix $\mathbf{S}^{-1}$.

In the general case we cannot make any assumptions about the features, and $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ acts as a transformation that decorrelates, centres and normalises the features.

Bivariate linear regression *

First, we derive the basic expressions.

$$
\begin{aligned}
\mathbf{X}^{\mathrm{T}}\mathbf{X} &= \begin{pmatrix} x_{11} & \cdots & x_{n1} \\ x_{12} & \cdots & x_{n2} \end{pmatrix} \begin{pmatrix} x_{11} & x_{12} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{pmatrix} = n \begin{pmatrix} \sigma_{11} + \overline{x_1}^2 & \sigma_{12} + \overline{x_1}\,\overline{x_2} \\ \sigma_{12} + \overline{x_1}\,\overline{x_2} & \sigma_{22} + \overline{x_2}^2 \end{pmatrix} \\
(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} &= \frac{1}{nD} \begin{pmatrix} \sigma_{22} + \overline{x_2}^2 & -\sigma_{12} - \overline{x_1}\,\overline{x_2} \\ -\sigma_{12} - \overline{x_1}\,\overline{x_2} & \sigma_{11} + \overline{x_1}^2 \end{pmatrix} \\
D &= (\sigma_{11} + \overline{x_1}^2)(\sigma_{22} + \overline{x_2}^2) - (\sigma_{12} + \overline{x_1}\,\overline{x_2})^2 \\
\mathbf{X}^{\mathrm{T}}\mathbf{y} &= \begin{pmatrix} x_{11} & \cdots & x_{n1} \\ x_{12} & \cdots & x_{n2} \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = n \begin{pmatrix} \sigma_{1y} + \overline{x_1}\,\overline{y} \\ \sigma_{2y} + \overline{x_2}\,\overline{y} \end{pmatrix}
\end{aligned}
$$

Bivariate linear regression *

We now consider two special cases. The first is that $\mathbf{X}$ is in homogeneous coordinates, i.e., we are really dealing with a univariate problem. In that case we have $x_{i1} = 1$ for $1 \leq i \leq n$; $\overline{x_1} = 1$; and $\sigma_{11} = \sigma_{12} = \sigma_{1y} = 0$. We then obtain (we write x instead of x_2, σ_{xx} instead of σ_{22} and σ_{xy} instead of σ_{2y}):

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \frac{1}{n\sigma_{xx}} \begin{pmatrix} \sigma_{xx} + \overline{x}^2 & -\overline{x} \\ -\overline{x} & 1 \end{pmatrix}$$

$$\mathbf{X}^{\mathrm{T}}\mathbf{y} = n \begin{pmatrix} \overline{y} \\ \sigma_{xy} + \overline{x}\,\overline{y} \end{pmatrix}$$

$$\hat{\mathbf{w}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \frac{1}{\sigma_{xx}} \begin{pmatrix} \sigma_{xx}\overline{y} - \sigma_{xy}\overline{x} \\ \sigma_{xy} \end{pmatrix}$$

This is the same result as obtained for the univariate case.

Bivariate linear regression *

The second special case we consider is where we assume x_1, x_2 and y to be *zero-centred*, which means that the intercept is zero and $\mathbf{w}$ contains the two regression coefficients. In this case we obtain

$$
\begin{aligned}
(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} &= \frac{1}{n(\sigma_{11}\sigma_{22} - \sigma_{12}^2)} \begin{pmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{12} & \sigma_{11} \end{pmatrix} \\
\mathbf{X}^{\mathrm{T}}\mathbf{y} &= n \begin{pmatrix} \sigma_{1y} \\ \sigma_{2y} \end{pmatrix} \\
\hat{\mathbf{w}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} &= \frac{1}{(\sigma_{11}\sigma_{22} - \sigma_{12}^2)} \begin{pmatrix} \sigma_{22}\sigma_{1y} - \sigma_{12}\sigma_{2y} \\ \sigma_{11}\sigma_{2y} - \sigma_{12}\sigma_{1y} \end{pmatrix}
\end{aligned}
$$

The last expression shows, e.g., that the regression coefficient for x_1 may be non-zero even if x_1 doesn't correlate with the target variable ($\sigma_{1y} = 0$), on account of the correlation between x_1 and x_2 ($\sigma_{12} \neq 0$).

Regularisation

Parameter Estimation by Optimization I

Regularisation is a general method to avoid overfitting by applying additional constraints to the weight vector. A common approach is to make sure the weights are, on average, small in magnitude: this is referred to as *shrinkage*.

Recall the setting for regression in terms of cost minimization.

- Can add penalty terms to a *cost* function, forcing coefficients to shrink to zero

$Y = a + b_1 X$ $\quad$ $Y = a + b_1 X + b_3 X^2$ $\quad$ $Y = a + b_1 X + b_3 X^2 + b_3 X^3 + b_3 X^3 + b_4 X^4$

$$Y = f_{\theta_0,\theta_1,\ldots,\theta_n}(X_1, X_2, \ldots, X_n) = f_\theta(\mathbf{X})$$

Parameter Estimation by Optimization II

- MSE as a cost function, given data $(\mathbf{x_1}, y_1), \ldots, (\mathbf{x_n}, y_n)$

$$Cost(\theta) = \frac{1}{n} \sum_i (f_\theta(\mathbf{x_i}) - y_i)^2$$

and with a penalty function:

$$Cost(\theta) = \frac{1}{n} \sum_i (f_\theta(\mathbf{x_i}) - y_i)^2 + \frac{1}{n} \lambda \sum_{i=1}^{n} \theta_i$$

- Parameter estimation by optimisation will attempt to values for $\theta_0, \theta_1, \ldots, \theta_n$ s.t. $Cost(\theta)$ is a minimum
- It will be easier to take the $\frac{1}{n}$ term as $\frac{1}{2n}$, which will not affect the minimisation

Parameter Estimation by Optimization III

- Using gradient descent with the penalty function will do two things: (a) we will move each θ_j in a direction that minimises the cost; and (b) each value of θ_j will also get "shrunk" on each iteration by multiplying the old value by an amount < 1

$$\theta_j^{(i+1)} = \alpha \theta_j^{(i)} - \eta \nabla_{\theta_j}$$

where $\alpha < 1$

Regularised regression

The multivariate least-squares regression problem can be written as an optimisation problem:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \; (\mathbf{y} - \mathbf{X}\mathbf{w})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\mathbf{w})$$

The regularised version of this optimisation is then as follows:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \; (\mathbf{y} - \mathbf{X}\mathbf{w})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda||\mathbf{w}||^2$$

where $||\mathbf{w}||^2 = \sum_i w_i^2$ is the squared norm of the vector $\mathbf{w}$, or, equivalently, the dot product $\mathbf{w}^{\mathrm{T}}\mathbf{w}$; λ is a scalar determining the amount of regularisation.

Regularised regression

This regularised problem still has a closed-form solution:

$$\hat{\mathbf{w}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda \mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

where $\mathbf{I}$ denotes the identity matrix. Regularisation amounts to adding λ to the diagonal of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, a well-known trick to improve the numerical stability of matrix inversion. This form of least-squares regression is known as *ridge regression*.

An interesting alternative form of regularised regression is provided by the *lasso*, which stands for 'least absolute shrinkage and selection operator'. It replaces the ridge regularisation term $\sum_i w_i^2$ with the sum of absolute weights $\sum_i |w_i|$. The result is that some weights are shrunk, but others are set to 0, and so the lasso regression favours *sparse solutions*.

Some further issues in learning linear regression models

What do the Coefficients b_i Mean?

- Consider the two equations:

$$\hat{Y} = a + bX$$

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$$

- b: change in Y that accompanies a unit change in X
- b_1: change in Y that accompanies a unit change in X_1 *provided* X_2 *remains constant*
- More generally, b_i $(i > 0)$ is the change in Y that accompanies a unit change in X_i provided all other X's are constant
- So: if all relevant variables are included, then we can assess the effect of each one in a controlled manner

Categoric Variables: X's I

- "Indicator" variables are those that take on the values 0 or 1
- They are used to include the effects of categoric variables. For example, if D is a variable that takes the value 1 if a patient takes a drug and 0 if the patient does not. Suppose you want to know the effect of drug D on blood pressure Y keeping age (X) constant

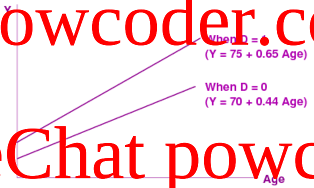
$$\hat{Y} = 70 + 5D + 0.44X$$

So, taking the drug (a unit change in D) makes a difference of 5 units, provided age is held constant

Categoric Variables: X's II

- How do we capture any interaction effect between age and drug intake? Introduce a new indicator variable $DX = D \times X$

$$\hat{Y} = 70 + 5D + 0.44X + 0.21DX$$



Categoric Values: Y values

- Sometimes, Y values are simply one of two values (let's call them 0 and 1)
- We can't use the regression model as we described earlier, in which the Y's can take any real value
- But, we can define a new linear regression model in which predicts not the value of Y, but what are called the *log odds* of Y:

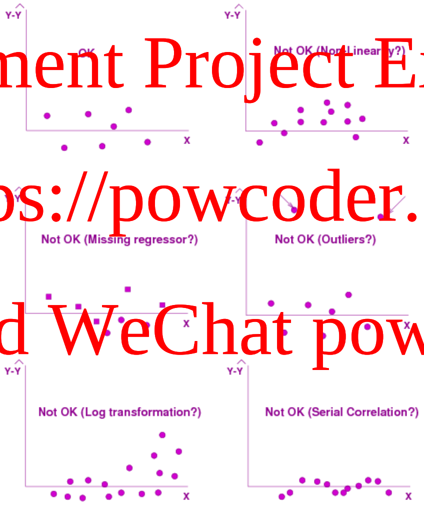
$$\log \text{odds } Y \; = \; Odds \; = \; b_0 + b_1 X_1 + \cdots + b_n X_n$$

- Once $Odds$ are estimated, they can be used to calculate the probability of Y:

$$Pr(Y = 1) = \frac{e^{Odds}}{(1 + e^{Odds})}$$

  We can then use the value of $Pr(Y = 1)$ to decide if $Y = 1$
- This procedure is called *logistic regression* (we'll see this again)

Is the Model Appropriate ? * I



Is the Model Appropriate ? * II

- The residuals from the regression line can be calculated numerically, along with their mean, variance and standard deviation. It can be shown that the residual standard deviation is related to the standard deviation of the Y values in the following manner:

$$rsd = s_y\sqrt{1-r^2}$$

- This helps us understand how much the regression line helped reduce the scatter of the y values (s_y gives a measure of the scatter of y values about the mean $\bar{y}$, and rsd gives a measure of the scatter of y values about the regression line)
- This also gives you another way of understanding the correlation coefficient. With $r = 0.9$, the scatter about the regression line is still almost 45% of the original scatter about the mean

Is the Model Appropriate ? * III

- If there is no systematic pattern to the residuals—that is, there are approximately half of them that are positive and half that are negative, then the line is a good fit

- It should also be the case that there should be no pattern to the residual scatter all along the line. If the average size of the residuals varies along the line (this condition is called *heteroscedasticity*) then the relationship is probably more complex than a straight line

- Residuals from a well-fitting line should show an approximate symmetric, bell-shaped frequency distribution with a mean of 0

Non-linear Relationships

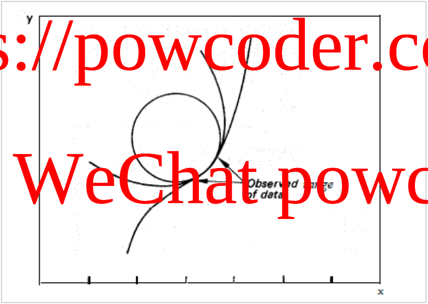
- Sometimes, the linear model may be inappropriate
- Some non-linear relationships can be captured in a linear model by a transformation ("trick"). For example, the curved model $\hat{Y} = b_0 + b_1 X_1 + b_2 X_1^2$ can be transformed by $X_2 = X_1^2$ into a linear model. This works for polynomial relationships.
- Some other non-linear relationships may require more complicated transformations. For example, the relationship is $Y = b_0 X_1^{b_1} X_2^{b_2}$ can be transformed into the linear relationship

$$\log(Y) = \log b_0 + b_1 \log X_1 + b_2 \log X_2$$

- Other relationships cannot be transformed quite so easily, and will require full non-linear estimation (in subsequent topics in the ML course we will find out more about these)

Non-Linear Relationships (contd.)

- Main difficulty with non-linear relationships is choice of function
  - How to learn ?
  - Can use a form of gradient descent to estimate the parameters
- After a point, almost any sufficiently complex mathematical function will do the job in a sufficiently small range



- Some kind of prior knowledge or theory is the only way to help here.
  - Otherwise, it becomes a process of trial-and-error, in which case, beware of conclusions that can be drawn

Model Selection

- Suppose there are a lot of variables X_i, some of which may be representing products, powers, *etc*.
- Taking all the X_i will lead to an overly complex model. There are 3 ways to reduce complexity:

1. Subset-selection, by search over subset lattice. Each subset results in a new model, and the problem is one of model-selection
2. Shrinkage, or *regularization* of coefficients to zero, by optimization. There is a single model, and unimportant variables have near-zero coefficients.

3. Dimensionality-reduction, by projecting points into a lower dimensional space (this is different to subset-selection, and we will look at it later)

Model Selection as Search I

- The subsets of the set of possible variables form a lattice with $S_1 \cap S_2$ as the g.l.b. or meet and $S_1 \cup S_2$ as the l.u.b. or join
- Each subset refers to a model, and a pair of subsets are connected if they differ by just 1 element
- A lattice is a graph, and we know how to search a graph
  - A^*, greedy, randomised *etc.*
  - "Cost" of node in the graph: MSE of the model. The parameters (coefficients) of the model can be found
- Historically, model-selection for regression has been done using "forward-selection", "backward-elimination", or "stepwise" methods
  - These are greedy search techniques that either: (a) start at the top of the subset lattice, and add variables; (b) start at the bottom of the subset lattice and remove variables; or (c) start at some interior point and proceed by adding or removing single variables (examining nodes connected to the node above or below)

Model Selection as Search II

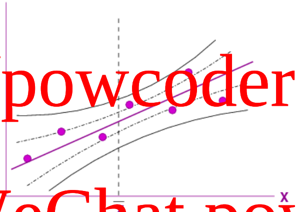
- Greedy selection done on the basis of calculating the *coefficient of determination* (often denoted by R^2) which denotes the proportion of total variation in the dependent variable Y that is explained by the model

- Given a model formed with a subset of variables X, it is possible to compute the observed change in R^2 due to the addition or deletion of some variable x
- This is used to select greedily the next best move in the graph-search

To set other *hyper-parameters*, such as shrinkage parameter λ, can use grid search

Prediction I

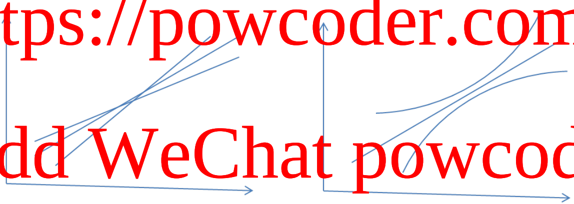
- It is possible to quantify what happens if the regression line is used for prediction:



- The intuition is this:
  - Recall the regression line goes through the mean $(\overline{X}, \overline{Y})$

Prediction II

- If the X_i are slightly different, then the mean is not going to change much. So, the regression line stays somewhat "fixed" at $(\overline{X}, \overline{Y})$ but with a different slope
- With each different sample of the X_i we will get a slightly different regression line
- The variation in Y values is greater further we move from $(\overline{X}, \overline{Y})$



- MORAL: Be careful, when predicting far away from the centre value
- ANOTHER MORAL: The model only works under the approximately the same conditions that held when collecting the data

Local (nearest-neighbour) regression

Local learning

- Related to the simplest form of learning: rote learning or memorization
- Training instances are searched for instance that **most closely resembles** *query* or test instance
- The *instances* themselves represent the knowledge
- Called: *nearest-neighbour*, *instance-based*, *memory-based* or *case-based* learning; all forms of *local learning*
- The *similarity* or *distance* function defines "learning", i.e., how to go beyond simple memorization
- Intuition — classify an instance similarly to examples "close by" — neighbours or *exemplars*
- A form of *lazy* learning – don't need to build a model!

Nearest neighbour for numeric prediction

Store all training examples $\langle x_i, f(x_i) \rangle$

Nearest neighbour:

- Given query instance x_q,
- first locate nearest training example x_n,
- then estimate $\hat{y} = \hat{f}(x_q) = f(x_n)$
- k-Nearest neighbour:
- Given x_q, take mean of f values of k nearest neighbours

$$\hat{y} = \hat{f}(x_q) = \frac{\sum_{i=1}^{k} f(x_i)}{k}$$

Distance function

The distance function defines what is "learned", i.e., predicted.
Instance x_i is described by an m vector of feature values:

$$\langle x_{i1}, x_{i2}, \ldots x_{im} \rangle$$

where x_{ik} denotes the value of the kth feature of x_i.
Most commonly used distance function is *Euclidean* distance, where the distance between two instances x_i and x_j is defined to be:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$$

Local regression

Use kNN to form a local approximation to f for each query point x_q using a linear function of the form

$$\hat{f}(x) = b_0 + b_1 x_1 + \ldots + b_m x_m$$

where x_i denotes the value of the ith feature of instance x.

Where does this linear regression model come from ?

- fit linear function to k nearest neighbours
- or quadratic or higher-order polynomial ...
- produces "piecewise approximation" to f

Summary

Summary

- Linear models give us a glimpse into many aspects of Machine Learning

  Terminology. Training data, test data, resubstitution error, prediction error.

  Conceptual. Learning as search, learning as optimisation, assumptions underlying a technique

  Implementation. Approximate alternatives to analytical methods

  Application. Overfitting, problems of prediction

  Each of these aspects will have counterparts in other kinds of machine learning

- Linear models are one way to predict numerical quantities
  - Ordinal regression: predicting ranks (not in the lectures)
  - Neural networks: non-linear regression models (later)
  - Regression trees: piecewise regression models (later)
  - Class-probability trees: predicting probabilities (later)
  - Model trees: piecewise non-linear models (later)