Supervised Learning – Regression

Last revision: 7 Mar 201

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.c

Material derived f
“Machine Learn
Cambridge Univ
http://cs.br

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 a

- the s
- how l
- fitting linear regression by least squares error cri
- non-linear regression via linear-in-the-para

- parameter estimation for regression
- local (nearest-neighbour) regression

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

Introduction

In the intro

classifica ata

instance

. . . 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.

Introduction

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

```
PRP =
   - 5
```



```
     + 0.006 MMAX
     + 0.630 CACH
     - 0.270 CHMIN
     + 1.46 CHMAX
```

Introduction

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

represented in a formal hypothesis language (trees, r

Introduction

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

represented as mathematical models (linear equati

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

Introduction

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

- line
  the le

- line
  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
- mul
  pre

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

Regression

We will look at the simplest model for numerical prediction:

a *reg*

Outcom

Note: the term *regression* is overloaded – it can re

- the process of determining the weights for the reg
- the regression equation itself.

Linear Regression

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
- Lin s

- Weights are calculated from the training data
- **Predicted** value for first training instance

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

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

Squared e

$$y \quad - \quad b_i x$$

$$j{=}1 \qquad\qquad i{=}0$$

Coefficients can be derived using standard matrix ope

Can be done if there are more instances than attributes (

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.

Probability vs Statistics: The Difference

- **Probability** versus **Statistics**
- Pro
  - al sense
- Statistics: reasons from samples to population
  - This is inductive reasoning, and is usually nse of the word)

Statistical Analyses

- Statistical analyses usually involve one of 3 things:
  1.
  2.
  3.

- Sta
  1. What is the question to be answered?
  2. Can it be quantitative (i.e. can we make measu
  3. How do we collect data?
  4. What can the data tell us?

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 tim
- For p
  mea
  idea

- *Sampling* is a way to draw conclusions about th
  having to measure all of the population. The concl
  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 c
- The calc conclusions.)
- The chance of obtaining an unrepresentative s 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
- Dra
  ele

Usually, t
n numbers that are approximately random.

In addition, the computer will use a mathematical relat
elements of the population and the set of numbers. Inver
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 mo dist er nu distribution
- For example, take a 2-stage sampling procedur are grouped according to size and the probabilit households is higher. A household is selected and 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 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 *con*

- So corr m is a very good estimate of the population mean μ. But this is not always the case. For example, the range of a sampl under-estimates the range of the population

- We will have to clarify what is meant by a "good esti 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 e
  eac
  mea
- Such an estimator is said to be *statis*

Sample Estimates of the Mean and the Spread I

Mean. This is calculated as follows.

Find the total T of N observations. Estimate the

ed

by "normal" distribution)

- A simple mathematical expr

$m = \frac{1}{N}\sum_i x_i$, where the ob $\dots x_n$

- If we can group the data so that the 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

y.

value of observations modelled by some theoretical probability distribution fun

similar counting method for c
random variables modelled u
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

Variance. This is calculated as follows:

- Calculate the total T a ... N observations. The estimate o $s = \sqrt{\frac{1}{N-1}\sum_i (x_i - m)^2}$
- 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

$$s = \sqrt{\frac{1}{N-1} \sum_i (x_i - m)^2 f_i}$$

X

$$Var(X) = \quad {}^2$$

$=$

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

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 c

  the

  valu he true value of the parameter θ. That is:

$$\text{MSE} = \text{Avg. val}$$

- 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

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

  the lowest possible value of MSE is 0
- In ge ... $0.$
  Sa

  esti ... o, given
  an estimator with bias b, we can calculate t
  variance of the estimator using the CR bound (sa :

$$\mathrm{MSE} \geq v$$

  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

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

Imagine t s of the
same size d
based on th
Then the

$$\text{MSE} = E[\hat{y} - f(x)]^2$$
$$= E[\hat{y} - E(\hat{y})]^2 + [$$

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 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
  - t
  - x and y
- Only appropriate when x and y are rou (doesn't work well when the association is curve
- The formula for computing correlation betwe

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

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:

- Wh

  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 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:

- Computers generally use a short-cut formula:

$$r = \frac{\sum_i x_i y_i}{n}$$

- 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 obs
(a "N

0;
and (
- Then:
  - The sampling distribution of the correlation c r varies from sample to sample) is also approxim

according to the Normal distribution with me ror (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

- $0.1,$

- with correlation 0.3, with a 95% confidenc 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
  give y value

  - different relationships
  - It is possible for two datasets to have different co
    same relationship

- MORAL: Do not use correlations to compare da
  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

- 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
  ma
  reas

- Remember, the correlation coefficient can tell
  such a relationship

- In real life, even if such a relationship held, it will be unr
  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?

Linear Relationship Between 2 Variables I

- GOAL: fit a line whose equation is of the form $Y = a + bX$
- 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:

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

whe

$\sum_i$

- This can be simplified to:

$$b = \sum(xy)$$

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 cha
- If the v wer nclude the change in X and any other confounding variables that may have changed as a result of changing X by example, that a change of X by 1 unit ca Y
- $b = 0$ means there is no linear relationship bet nd 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 (sta
- Wh

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 ass
  -
  - s on a
  - The Y_i are independent
- In standard terminology, the Y_i are ide ent (i.i.d.) random variables with mean μ 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 abo
- It ca
  esti
  lowest variance
- The proof of this is called the *Gauss-*
  Gauss-Markov theorem makes the following
  1. The expected (average) values of residuals is 0 ($_i$
  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 hav
- The resi distribution, with mean 0
  - In this case, minimising least-squares is equiv probability of the Y_i, given the X_i (to *maximum likelihood estimation*)

  - More on this in a later lecture

Univariate linear regression

Example:

Suppose
and weig
(h_i, w_i),

Univariate linear regression assumes a linear equati ith
parameters a and b chosen such that the sum of sq
$\sum_{i=1}^{n}(w_i - (a + bh_i))^2$ is minimised.

Univariate linear regression

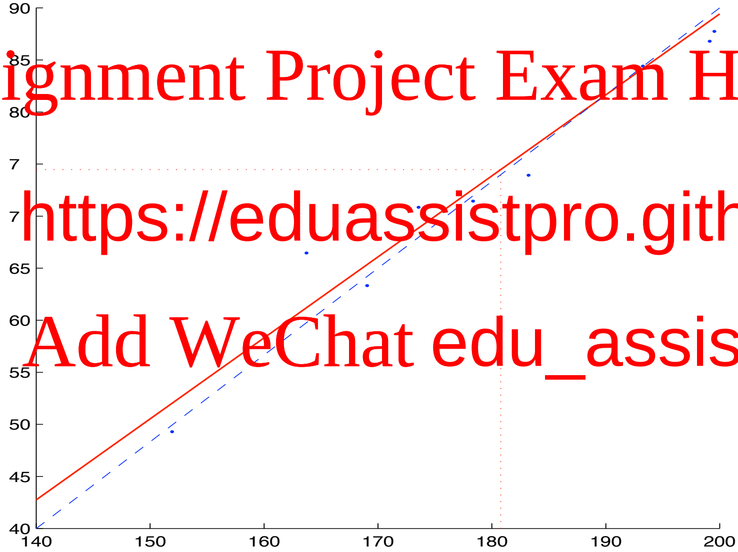
In order to find the parameters we take partial derivatives, set the partial derivatives to 0 and solve for a and b:

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

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

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 soli

measure

ody

height (on t

the average height $\overline{h} = 181$ and the average w

regression coefficient $\hat{b} = 0.78$. The measure

adding normally distributed noise with mean 0 and var

model indicated by the blue dashed line (b

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:

(Here we u xx x

This can be understood by noting that the covariance is of x times units of y (e.g., metres times kilogr 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 co
intercept
deviatio

So we could *zero-centre* the x-values by subt
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 $\bar{x}$ with $\overline{x'} = \bar{x}/\sigma_{xx}$, then we have that $\hat{b} = \frac{1}{n}\sum_{i=1}^{n}(x'_i - \overline{x'})(y_i - \bar{y}) = \sigma_{x'y}$.

In other wo 's variance, we can tak variable a

This demonstrates that univariate linear regressio consisting of two steps:

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

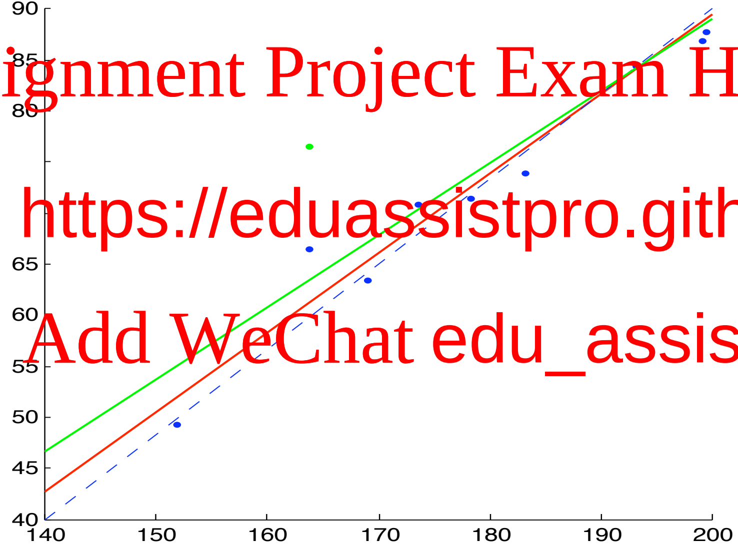
Linear regression: intuitions

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

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

While this property is intuitively appealing, it is worth k 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 t
values fro
10 kg. The di
least-squares regression line.

Specifically, we see that one of the blue points got moved u
the green point, changing the red regression line to the gr

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 diff
- A nu

  step

  stopping when we reach a minimum
- Recall that at a point the gradient vector points in th greatest increase of a function. So, the opposite di gradient vector gives the direction of greatest de

  - $b_{i+1} = b_i$ - $\eta \times g_b$
  - $a_{i+1} = a_i$ - $\eta \times g_a$
  - Stop when $b_{i+1} \approx b_i$ and $a_{i+1} \approx a_i$
- More on this in a later lecture

Many variables

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

gen
carcinogenicity to be related to some surrogate v
example)

- Including more variables can give a narrower co the prediction being made

Multivariate linear model

- The Y_i are identically distributed independent variables with mean $\mu = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \quad + \beta_n X_n$ and variance σ^2
- Or: dependent erro

- As be equation $\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \cdots$
- With many variables, the regression equation a b_i are expressed better using a matrix represen equations.

Multivariate linear regression *

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

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

Assumin

$\mu_j = 0$ a

es

(times n

We can normalise the features by means of a

:

diagonal matrix with diagonal entries $1/n$

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 $\mathbf{S}$:

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

Let us try to u

- Ass ... rix $\mathbf{\Sigma}$ is dia
- Assuming the features are zero-centred, $^{\mathrm{T}}$ ... nal with entries $n\sigma_{jj}$.
- In other words, assuming zero-centred and unc ... $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ reduces to our scaling matrix $\mathbf{S}^{-}$.

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.

$$\mathbf{X}^{\mathrm{T}} \quad x_{11} \quad x_{12} \quad \ldots \quad {}_{2} + \overline{x_1}\,\overline{x_2} \quad {}_{22} + \overline{x_2}^2 \Big)$$

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X}) \quad \overline{nD} \quad -\sigma_{12} - \overline{x_1}\,\overline{x_2} \quad \sigma_{11} + \overline{x_1}$$

$$D = (\sigma_{11} + \overline{x_1}^2)(\sigma_{22} + \overline{x_2}^2) - (\sigma_{12} + \overline{x_1}\,\overline{x}$$

$$\mathbf{X}^{\mathrm{T}}\mathbf{y} = \begin{pmatrix} x_{11} & \cdots & x_{n1} \\ x_{12} & \cdots & x_{n2} \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = n$$

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 [illegible]

$$[illegible] \frac{[illegible]}{n\sigma_{xx}} \left([illegible] \quad \overline{x} \quad 1 \right)$$

$$\mathbf{X}^{\mathrm{T}}\mathbf{y} = n \left([illegible] \right)$$

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

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

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

$$\mathbf{X}^{\mathrm{T}}\mathbf{y} = n\begin{pmatrix}\sigma_{1y}\\ \sigma_{2y}\end{pmatrix}$$

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

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 ($\sigma_{12} \neq 0$).

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 s

- Can s to shrink to zero

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

Parameter Estimation by Optimization II

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

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

and

$$Cost(\theta) = \frac{1}{n}\sum(f_\theta(\mathbf{x_i} \qquad {}^2 \quad \frac{1}{\;} \qquad n$$

- Parameter estimation by optimisation will att $\theta_0, \theta_1, \ldots, \theta_n$ 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) w

  (b) e

  mul

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

where $\alpha < 1$

Regularised regression

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

The regul

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

where $||\mathbf{w}||^2 = \sum_i w_i^2$ is the squared norm of the v
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:

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

where $\mathbf{I}$ λ
to the diag

stability o
as *ridge regression*.

An interesting alternative form of regularised regres
lasso, which stands for 'least absolute shrinkage and se
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*.

What do the Coefficients b_i Mean?

- Consider the two equations:

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

- b: cha
- b_1: change in Y that accompanies a unit ch *ed* X_2 *remains constant*

- More generally, b_i $(i > 0)$ is the change in nit 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
- The
  exa akes a
  dru
  effe stant

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

So, taking the drug (a unit change in
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
not

$$\log \text{odds}\; Y = Odds = b + b\ X + \qquad + b\ X$$

- Once $Odds$ are estimated, they can be used to c
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:

- This helps us understand how much the regression line helped reduce the scatter of the y values (s_y gives a m [illegible] y values about the mean $\bar{y}$, and rsd giv [illegible] 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 neg

- It sho resi varies along the line (this condition is called) then the relationship is probably more complex than

- Residuals from a well-fitting line should show an a symmetric, bell-shaped frequency distribut 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 $\hat{Y}$ into a linear mo

- So transformations. For example, the relationship is $Y = b \ X^{b_1} X_2^{b_2}$ can be transformed into the linear relationship

$$\log(Y) = \log b_0 + b_1 \mathrm{l}$$

- 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 d

- 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*.
- Ta
  way

  1. new model, and the problem is one of model-selection
  2. Shrinkage, or *regularization* of coeffici
     There is a single model, and unimportant varia
     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 the
- A lat
  -
  -
    (coefficients) of the model can be found
- Historically, model-selection for regression "forward-selection", "backward-elimin
  - 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 ... the
- ... le to ... n of some variable x
- This is used to select greedily the next best move in t

To set other *hyper-parameters*, such as shrinka ... e 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
- $(\overline{X}, \overline{Y})$

- 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 learning

- Related to the simplest form of learning: rote learning or memorization

- Training instances are searched for instance that **most closely res**

- The

- Call or *case-based* learning; all forms of *local l*

- The *similarity* or *distance* function defi 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:

- Giv
- first l
- the

- k-Nearest neighbour:
- Given x_q, take mean of f values of k

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

Distance function

The distance function defines what is "learned", i.e., predicted.

Instance x_i is described by an m vector of feature values:

where x_i

Most com here the distance between two instances x_i and x_j

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

Local regression

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

where x_i

Where does this linear regression model come from ?

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

Summary

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

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

  Co

  Imp

  Application. Overfitting, problems of prediction

  Each of these aspects will have counterparts in ot

  machine learning

- Linear models are one way to predict numerical q
  - 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)