Linear Regression

MLAI Lecture 4

Neil D. Lawrence

Department of Computer Science
Sheffield University

2nd October 2012

Outline

Regression Examples

Overdetermined Systems

Gaussian Density Reminder

Linear Regression

Regression Examples

- Predict a real value, t_i given some inputs $\mathbf{x}_i$.
- Predict quality of meat given spectral measurements (Tecator data).
- Radiocarbon dating, the C14 calibration curve: predict age given quantity of C14 isotope.
- Predict quality of different Go or Backgammon moves given expert rated training data.

Outline

Regression Examples

Overdetermined Systems

Gaussian Density Reminder

Linear Regression

Two Simultaneous Equations

A system of two simultaneous equations with two unknowns.

$$
\begin{aligned}
t_1 &= mx_1 + c \\
t_2 &= mx_2 + c
\end{aligned}
$$

Two Simultaneous Equations

A system of two simultaneous equations with two unknowns.

$$t_1 - t_2 = m(x_1 - x_2)$$

Two Simultaneous Equations

A system of two simultaneous equations with two unknowns.

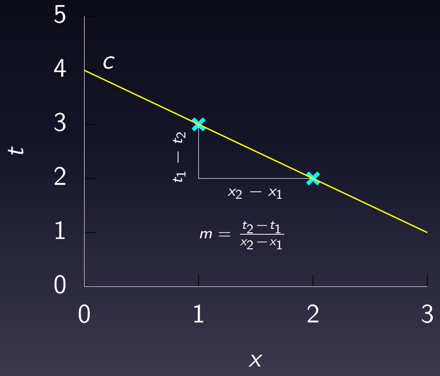
$$\frac{t_1 - t_2}{x_1 - x_2} = m$$

Two Simultaneous Equations

A system of two simultaneous equations with two unknowns.

$$m = \frac{t_2 - t_1}{x_2 - x_1}$$

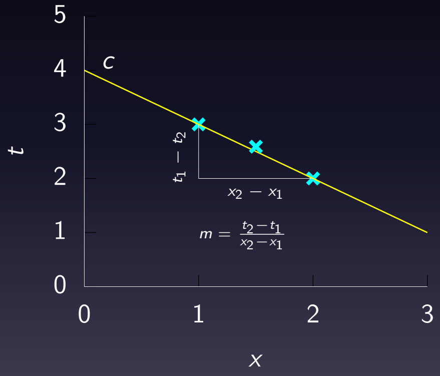
$$c = t_1 - m x_1$$



Two Simultaneous Equations

How do we deal with three simultaneous equations with only two unknowns?

$$t_1 = mx_1 + c$$
$$t_2 = mx_2 + c$$
$$t_3 = mx_3 + c$$



Overdetermined System

- With two unknowns and two observations:

$$t_1 = mx_1 + c$$
$$t_2 = mx_2 + c$$

- Additional observation leads to *overdetermined* system.

$$t_3 = mx_3 + c$$

- This problem is solved through a noise model $\epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$

$$t_1 = mx_1 + c + \epsilon_1$$
$$t_2 = mx_2 + c + \epsilon_2$$
$$t_3 = mx_3 + c + \epsilon_3$$

Overdetermined System

- With two unknowns and two observations:

$$
\begin{aligned}
t_1 &= mx_1 + c \\
t_2 &= mx_2 + c
\end{aligned}
$$

- Additional observation leads to *overdetermined* system.

$$t_3 = mx_3 + c$$

- This problem is solved through a noise model $\epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$

$$
\begin{aligned}
t_1 &= mx_1 + c + \epsilon_1 \\
t_2 &= mx_2 + c + \epsilon_2 \\
t_3 &= mx_3 + c + \epsilon_3
\end{aligned}
$$

Overdetermined System

- With two unknowns and two observations:

$$
\begin{aligned}
t_1 &= mx_1 + c \\
t_2 &= mx_2 + c
\end{aligned}
$$

- Additional observation leads to *overdetermined* system.

$$
t_3 = mx_3 + c
$$

- This problem is solved through a noise model $\epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$

$$
\begin{aligned}
t_1 &= mx_1 + c + \epsilon_1 \\
t_2 &= mx_2 + c + \epsilon_2 \\
t_3 &= mx_3 + c + \epsilon_3
\end{aligned}
$$

Noise Models

- We aren't modeling entire system.
- Noise model gives mismatch between model and data.
- Gaussian model justified by appeal to central limit theorem.
- Other models also possible (Student-t for heavy tails).
- Maximum likelihood with Gaussian noise leads to *least squares*.

Outline

Regression Examples

Overdetermined Systems

Gaussian Density Reminder

Linear Regression

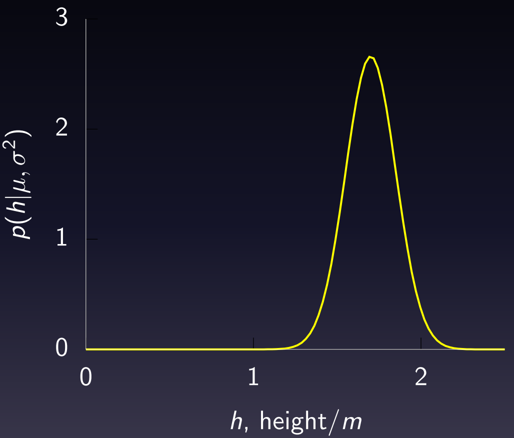
The Gaussian Density

- Perhaps the most common probability density.

$$p(t|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right)$$
$$= \mathcal{N}\left(t|\mu,\sigma^2\right)$$

- The Gaussian density.

Gaussian Density



The Gaussian PDF with $\mu = 1.7$ and variance $\sigma^2 = 0.0225$. Mean shown as red line. It could represent the heights of a population of students.

Gaussian Density

$$\mathcal{N}\left(t|\mu,\sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right)$$

Outline

Regression Examples

Overdetermined Systems

Gaussian Density Reminder

Linear Regression

A Probabilistic Process

- Set the mean of Gaussian to be a function.

$$p\left(t_i|x_i\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(t_i - y\left(x_i\right)\right)^2}{2\sigma^2}\right).$$

- This gives us a ‘noisy function’.
- This is known as a process.

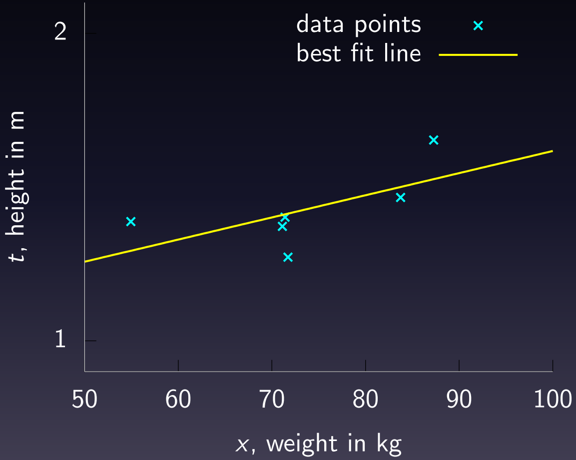
Height as a Function of Weight

- In the standard Gaussian, parametized by mean and variance.
- Make the mean a linear function of an *input*.
- This leads to a regression model.

$$
\begin{aligned}
t_i =& y\left(x_i\right) + \epsilon_i, \\
\epsilon_i \sim& \mathcal{N}\left(0, \sigma^2\right).
\end{aligned}
$$

- Assume t_i is height and x_i is weight.

Linear Function



A linear regression between height and weight.

- Likelihood of an individual data point

$$p\left(t_i|x_i, m, c\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(t_i - mx_i - c\right)^2}{2\sigma^2}\right).$$

- Parameters are gradient, m, offset, c of the function and noise variance σ^2.

Likelihood Function

- Assume samples are independent and identically distributed given the parameters (i.i.d.)
- Leads to the log likelihood

$$L(m, c, \sigma^2) = -\frac{N}{2} \log 2\pi - \frac{N}{2} \log \sigma^2 - \sum_{i=1}^{N} \frac{(t_i - mx_i - c)^2}{2\sigma^2}.$$

Error Function

- Negative log likelihood is the error function leading to an error function

$$E(m, c, \sigma^2) = \frac{N}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\sum_{i=1}^{N}(t_i - mx_i - c)^2 .$$

- Learning proceeds by minimizing this error function for the data set provided.

Connection: Sum of Squares Error

- Ignoring terms which don't depend on m and c gives

$$E(m, c) \propto \sum_{i=1}^{N} (t_i - y(x_i))^2$$

  where $y(x_i) = mx_i + c$.
- This is known as the *sum of squares* error function.
- Commonly used and is closely associated with the Gaussian likelihood.

Fixed Point Updates

Worked example.

$$c^* = \frac{\sum_{i=1}^{N} (t_i - m^* x_i)}{N},$$

$$m^* = \frac{\sum_{i=1}^{N} x_i (t_i - c^*)}{\sum_{i=1}^{N} x_i^2},$$

$$\sigma^{2\,*} = \frac{\sum_{i=1}^{N} (t_i - m^* x_i - c^*)^2}{N}$$

Multi-dimensional Inputs

- Multivariate functions involve more than one input.
- Height might be a function of weight and gender.
- There could be other contributory factors.
- Place these factors in a feature vector $\mathbf{x}_i$.
- Linear function is now defined as

$$y(\mathbf{x}_i) = \sum_{j=1}^{q} w_j x_{i,j} + c$$

Vector Notation

- Write in vector notation,

$$y(\mathbf{x}_i) = \mathbf{w}^\top \mathbf{x}_i + c$$

- Can absorb c into $\mathbf{w}$ by assuming extra input x_0 which is always 1.

$$y(\mathbf{x}_i) = \mathbf{w}^\top \mathbf{x}_i$$

Log Likelihood for Multivariate Regression

- The likelihood of a single data point is

$$p\left(t_i|x_i\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(t_i - \mathbf{w}^\top \mathbf{x}_i\right)^2}{2\sigma^2}\right).$$

- Leading to a log likelihood for the data set of

$$L(\mathbf{w}, \sigma^2) = -\frac{N}{2}\log\sigma^2 - \frac{N}{2}\log 2\pi - \frac{\sum_{i=1}^{N}\left(t_i - \mathbf{w}^\top \mathbf{x}_i\right)^2}{2\sigma^2}.$$

- And a corresponding error function of

$$E(\mathbf{w}, \sigma^2) = \frac{N}{2}\log\sigma^2 + \frac{\sum_{i=1}^{N}\left(t_i - \mathbf{w}^\top \mathbf{x}_i\right)^2}{2\sigma^2}.$$

Expand the Brackets

$$
\begin{aligned}
E(\mathbf{w}, \sigma^2) =& \frac{N}{2} \log \sigma^2 + \frac{1}{2\sigma^2} \sum_{i=1}^{N} t_i^2 - \frac{1}{\sigma^2} \sum_{i=1}^{N} t_i \mathbf{w}^\top \mathbf{x}_i \\
&+ \frac{1}{2\sigma^2} \sum_{i=1}^{N} \mathbf{w}^\top \mathbf{x}_i \mathbf{x}_i^\top \mathbf{w} + \text{const.} \\
=& \frac{N}{2} \log \sigma^2 + \frac{1}{2\sigma^2} \sum_{i=1}^{N} t_i^2 - \frac{1}{\sigma^2} \mathbf{w}^\top \sum_{i=1}^{N} \mathbf{x}_i t_i \\
&+ \frac{1}{2\sigma^2} \mathbf{w}^\top \left[\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^\top \right] \mathbf{w} + \text{const.}
\end{aligned}
$$

Multivariate Derivatives

- We will need some multivariate calculus.
- For now some simple multivariate differentiation:

$$\frac{\mathrm{d}\mathbf{a}^\top \mathbf{w}}{\mathrm{d}\mathbf{w}} = \mathbf{a}$$

and

$$\frac{\mathrm{d}\mathbf{w}^\top \mathbf{A}\mathbf{w}}{\mathrm{d}\mathbf{w}} = \left(\mathbf{A} + \mathbf{A}^\top\right)\mathbf{w}$$

or if $\mathbf{A}$ is symmetric ($i.e.$ $\mathbf{A} = \mathbf{A}^\top$)

$$\frac{\mathrm{d}\mathbf{w}^\top \mathbf{A}\mathbf{w}}{\mathrm{d}\mathbf{w}} = 2\mathbf{A}\mathbf{w}.$$

Differentiate

Differentiating with respect to the vector **w** we obtain

$$\frac{\partial L(\mathbf{w}, \beta)}{\partial \mathbf{w}} = \beta \sum_{i=1}^{N} \mathbf{x}_i t_i - \beta \left[\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^\top \right] \mathbf{w}$$

Leading to

$$\mathbf{w}^* = \left[\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^\top \right]^{-1} \sum_{i=1}^{N} \mathbf{x}_i t_i,$$

Rewrite in matrix notation:

$$\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^\top = \mathbf{X}^\top \mathbf{X}$$

$$\sum_{i=1}^{N} \mathbf{x}_i t_i = \mathbf{X}^\top \mathbf{t}$$

Update Equations

- Update for $\mathbf{w}^*$.

$$\mathbf{w}^* = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{t}$$

- The equation for ${\sigma^2}^*$ may also be found

$$ {\sigma^2}^* = \frac{\sum_{i=1}^{N} \left(t_i - {\mathbf{w}^*}^\top \mathbf{x}_i\right)^2}{N}.$$

Reading

- Section 1.2.5 of Bishop up to equation 1.65.
- Section 1.1 of Bishop as preparation for Friday.
- Section 1.1-1.2 of Rogers and Girolami for fitting linear models.
- Section 1.3 of Rogers and Girolami for Matrix & Vector Review.

References I

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, 2006. [Google Books] .

S. Rogers and M. Girolami. *A First Course in Machine Learning*. CRC Press, 2011. [Google Books] .