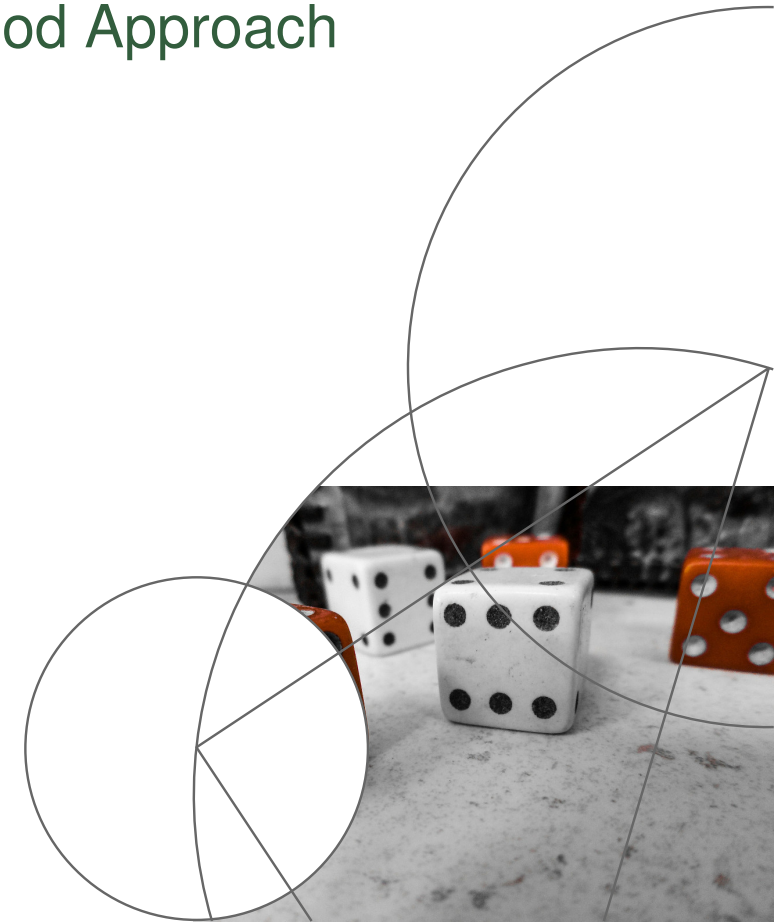
Linear Modelling: A Maximum Likelihood Approach

Modelling and Analysis of Data (MAD)

Kim Steenstrup Pedersen
Based on Fabian Gieseke's slides



December 2021

Outline

Outline

What we have seen so far . . .

1. Multivariate linear and non-linear regression & Regularisation
2. Basic probability theory and statistics
3. Advanced probability theory and statistics
4. Today: Maximum likelihood estimation

What we have seen so far . . .

1. Multivariate linear and non-linear regression & Regularisation
   (maybe the most prominent ML method; extensions: kernel methods!)
2. Basic probability theory and statistics
   (needed when defining probabilistic ML methods)
3. Advanced probability theory and statistics
   (needed when proving properties of ML methods)
4. Today: Maximum likelihood estimation
   (first step towards Bayesian learning)

The plan for the remainder of the course

6. Today: Maximum likelihood estimation
7. Bayesian Learning
8. Principal Components Analysis (PCA)
9. Classification (k-NN, SVMs, ...)
10. Sampling & Bayesian Inference (Sampling)
11. Unsupervised Learning & Clustering (k-means, . . .)

Outline

Random Variables and Probabilities

Remark: Today we will make use of random variables and probability theory as a modelling tool (not much though!). We will need a bit more for the next lecture(s). Remember what we talked about in the last lectures on probability theory and statistics . . .

Example

Bernoulli Distribution

Given a random variable X with two possible outcomes, 0 or 1, where the probability that it takes the value 1 is $q \in (0,1)$. Then, the Bernoulli distribution is given by

$$P(X = x) = q^x (1-q)^{1-x},$$

i.e., $P(X = 1) = q$ and $P(X = 0) = 1 - q$. We write $X \sim B(q)$.

Example

Bernoulli Distribution

Given a random variable X with two possible outcomes, 0 or 1, where the probability that it takes the value 1 is $q \in (0,1)$. Then, the Bernoulli distribution is given by

$$P(X = x) = q^x (1-q)^{1-x},$$

i.e., $P(X = 1) = q$ and $P(X = 0) = 1 - q$. We write $X \sim B(q)$.

Repeated coin toss: $X_i \sim B(p)$.
For instance $X_1 = 0$, $X_2 = 1$, $X_3 = 1$, $\ldots$, $X_N = 0$.

Random Variables and Probabilities

Comment 2.2 – Vector Random Variables (Rogers & Girolami)

It will often be necessary to define probability distributions over vectors. This is nothing more than a shorthand way of defining large joint distributions. For example, the values / samples that could be taken on by random variables $X_1, \ldots, X_N$ *can be expressed as the vector* $\mathbf{x} = [x_1, \ldots, x_N]^T$*. Using this shorthand:*

$$p(\mathbf{x}) = p(x_1, \ldots, x_N) = P(X_1 = x_1, \ldots, X_N = x_N)$$

Even though $\mathbf{x}$ *is a vector,* $p(\mathbf{x})$ *is a scalar quantity, just as* $P(X_1 = x_1, \ldots, X_N = x_N)$ *is.*

Maximum Likelihood Estimation?

Goal

Given some data $x_1, \ldots, x_N$ sampled from some random variable X with distribution p. How can I find a model distribution that describes the data distribution well?

Maximum Likelihood Estimation?

Goal

Given some data $x_1, \ldots, x_N$ sampled from some random variable X with distribution p. How can I find a model distribution that describes the data distribution well?

1. Without any assumptions: Tricky!

Maximum Likelihood Estimation?

Goal

Given some data $x_1, \ldots, x_N$ sampled from some random variable X with distribution p. How can I find a model distribution that describes the data distribution well?

1. Without any assumptions: Tricky!
2. In case we consider some specific distribution $p(x; \theta)$ that is parameterized by some $\theta \in \mathbb{R}$, it becomes easier. The distribution could also be described by multiple parameters, e.g., by a vector $\theta \in \mathbb{R}^k$.

Example: Maximum Likelihood Estimation

Task

Given some data $x_1, \ldots, x_N$ drawn (i.i.d.) from a Bernoulli distribution $B(q)$ with some unknown $q \in (0,1)$, i.e.:

$$P(X_n = x_n) = p(x_n) = q^{x_n}(1-q)^{1-x_n}$$

i.i.d. = independent and identically distributed
Notation: $p(x_n) = p(x_n; q) = p(x_n|q)$

Example: Maximum Likelihood Estimation

Task

Given some data $x_1, \ldots, x_N$ drawn (i.i.d.) from a Bernoulli distribution $B(q)$ with some unknown $q \in (0,1)$, i.e.:

$$P(X_n = x_n) = p(x_n) = q^{x_n}(1-q)^{1-x_n}$$

Question: How do you estimate the unknown parameter q?

i.i.d. = independent and identically distributed
Notation: $p(x_n) = p(x_n; q) = p(x_n|q)$

Example: Maximum Likelihood Estimation

Task

Given some data $x_1, \ldots, x_N$ drawn (i.i.d.) from a Bernoulli distribution $B(q)$ with some unknown $q \in (0,1)$, i.e.:

$$P(X_n = x_n) = p(x_n) = q^{x_n}(1-q)^{1-x_n}$$

Question: How do you estimate the unknown parameter q? One possible way to obtain an estimate is to compute the sample mean

$$\hat{q} = \frac{\sum_{n=1}^{N} x_n}{N}$$

i.i.d. = independent and identically distributed
Notation: $p(x_n) = p(x_n; q) = p(x_n|q)$

Example: Maximum Likelihood Estimation

Idea

Find $\hat{q} \in (0,1)$ that maximizes the probability of the observed data $x_1, \ldots, x_N$:

$$\underset{q}{\text{maximize}}\, p(x_1, \ldots, x_N; q) = \underset{q}{\text{maximize}} \prod_{n=1}^{N} p(x_n; q) = \underset{q}{\text{maximize}} \prod_{n=1}^{N} q^{x_n} (1-q)^{1-x_n}$$

Notice that $p(x_1, \ldots, x_N; q)$ is a function only of q, since the observed data $x_1, \ldots, x_N$ is fixed,

Example: Maximum Likelihood Estimation

Idea

Find $\hat{q} \in (0,1)$ that maximizes the probability of the observed data $x_1, \ldots, x_N$:

$$\underset{q}{\text{maximize}}\, p(x_1, \ldots, x_N; q) = \underset{q}{\text{maximize}} \prod_{n=1}^{N} p(x_n; q) = \underset{q}{\text{maximize}} \prod_{n=1}^{N} q^{x_n}(1-q)^{1-x_n}$$

Notice that $p(x_1, \ldots, x_N; q)$ is a function only of q, since the observed data $x_1, \ldots, x_N$ is fixed,

- Let's consider the log of the objective above, i.e., we want to maximize

$$l(q) = \log L(q) = \log\left(\prod_{n=1}^{N} q^{x_n}(1-q)^{1-x_n}\right) = \sum_{n=1}^{N} x_n \log q + (1-x_n)\log(1-q)$$

Example: Maximum Likelihood Estimation

Idea

Find $\hat{q} \in (0,1)$ that maximizes the probability of the observed data $x_1, \ldots, x_N$:

$$\underset{q}{\text{maximize}}\, p(x_1, \ldots, x_N; q) = \underset{q}{\text{maximize}} \prod_{n=1}^{N} p(x_n; q) = \underset{q}{\text{maximize}} \prod_{n=1}^{N} q^{x_n}(1-q)^{1-x_n}$$

Notice that $p(x_1, \ldots, x_N; q)$ is a function only of q, since the observed data $x_1, \ldots, x_N$ is fixed,

- Let's consider the log of the objective above, i.e., we want to maximize

$$l(q) = \log L(q) = \log\left(\prod_{n=1}^{N} q^{x_n}(1-q)^{1-x_n}\right) = \sum_{n=1}^{N} x_n \log q + (1-x_n)\log(1-q)$$

- Well, that's an optimization task with one variable. Pause the video and think about how to solve this!

Example: Maximum Likelihood Estimation

Optimization Problem

$$\underset{q}{\text{maximize}}\, l(q) = \underset{q}{\text{maximize}} \sum_{n=1}^{N} x_n \log q + (1 - x_n) \log(1 - q)$$

We have: $\frac{\partial l}{\partial q} = \sum_{n=1}^{N} \frac{x_n}{q} - \frac{1-x_n}{1-q}$. A necessary condition for an optimum is:

$$\frac{\partial l}{\partial q} = 0 \Leftrightarrow \qquad \sum_{n=1}^{N} \frac{x_n}{q} - \frac{1-x_n}{1-q} \qquad = 0$$

Example: Maximum Likelihood Estimation

Optimization Problem

$$\underset{q}{\text{maximize}}\, l(q) = \underset{q}{\text{maximize}} \sum_{n=1}^{N} x_n \log q + (1 - x_n) \log(1 - q)$$

We have: $\frac{\partial l}{\partial q} = \sum_{n=1}^{N} \frac{x_n}{q} - \frac{1-x_n}{1-q}$. A necessary condition for an optimum is:

$$\begin{aligned} \frac{\partial l}{\partial q} = 0 \Leftrightarrow \quad & \sum_{n=1}^{N} \frac{x_n}{q} - \frac{1-x_n}{1-q} = 0 \\ \Leftrightarrow \quad & \frac{1}{q} \sum_{n=1}^{N} x_n = \frac{1}{1-q} \sum_{n=1}^{N} (1 - x_n) \end{aligned}$$

Example: Maximum Likelihood Estimation

Optimization Problem

$$\underset{q}{\text{maximize}}\, l(q) = \underset{q}{\text{maximize}} \sum_{n=1}^{N} x_n \log q + (1 - x_n) \log(1 - q)$$

We have: $\frac{\partial l}{\partial q} = \sum_{n=1}^{N} \frac{x_n}{q} - \frac{1-x_n}{1-q}$. A necessary condition for an optimum is:

$$\begin{aligned}
\frac{\partial l}{\partial q} = 0 \Leftrightarrow \quad & \sum_{n=1}^{N} \frac{x_n}{q} - \frac{1-x_n}{1-q} = 0 \\
\Leftrightarrow \quad & \frac{1}{q} \sum_{n=1}^{N} x_n = \frac{1}{1-q} \sum_{n=1}^{N} (1 - x_n) \\
\Leftrightarrow \quad & \sum_{n=1}^{N} x_n - q \sum_{n=1}^{N} x_n = q \sum_{n=1}^{N} (1 - x_n)
\end{aligned}$$

Example: Maximum Likelihood Estimation

Optimization Problem

$$\underset{q}{\text{maximize}}\, l(q) = \underset{q}{\text{maximize}} \sum_{n=1}^{N} x_n \log q + (1 - x_n) \log(1 - q)$$

We have: $\frac{\partial l}{\partial q} = \sum_{n=1}^{N} \frac{x_n}{q} - \frac{1-x_n}{1-q}$. A necessary condition for an optimum is:

$$\begin{aligned}
\frac{\partial l}{\partial q} = 0 \Leftrightarrow \quad & \sum_{n=1}^{N} \frac{x_n}{q} - \frac{1-x_n}{1-q} && = 0 \\
\Leftrightarrow \quad & \frac{1}{q} \sum_{n=1}^{N} x_n && = \frac{1}{1-q} \sum_{n=1}^{N} (1 - x_n) \\
\Leftrightarrow \quad & \sum_{n=1}^{N} x_n - q \sum_{n=1}^{N} x_n && = q \sum_{n=1}^{N} (1 - x_n) \\
\Leftrightarrow \quad & \frac{\sum_{n=1}^{N} x_n}{N} && = q
\end{aligned}$$

Example: Maximum Likelihood Estimation

Optimization Problem

$$\underset{q}{\text{maximize}}\, l(q) = \underset{q}{\text{maximize}} \sum_{n=1}^{N} x_n \log q + (1 - x_n) \log(1 - q)$$

We have: $\frac{\partial l}{\partial q} = \sum_{n=1}^{N} \frac{x_n}{q} - \frac{1-x_n}{1-q}$. A necessary condition for an optimum is:

$$\begin{aligned}
\frac{\partial l}{\partial q} = 0 \Leftrightarrow \quad & \sum_{n=1}^{N} \frac{x_n}{q} - \frac{1-x_n}{1-q} = 0 \\
\Leftrightarrow \quad & \frac{1}{q} \sum_{n=1}^{N} x_n = \frac{1}{1-q} \sum_{n=1}^{N} (1 - x_n) \\
\Leftrightarrow \quad & \sum_{n=1}^{N} x_n - q \sum_{n=1}^{N} x_n = q \sum_{n=1}^{N} (1 - x_n) \\
\Leftrightarrow \quad & \frac{\sum_{n=1}^{N} x_n}{N} = q
\end{aligned}$$

Since $\frac{\partial^2 l}{\partial q \partial q} = -\frac{\sum_{n=1}^{N} x_n}{q^2} - \frac{\sum_{n=1}^{N} (1-x_n)}{(1-q)^2} < 0$, the estimate $\hat{q} = \frac{\sum_{n=1}^{N} x_n}{N}$ is a global maximum.

(since $0 < \sum_{n=1}^{N} x_n < N$ and $0 < q < 1$)

General Case

Maximum Likelihood Estimation

Given some dataset $x_1, \ldots, x_N$ drawn independently from some distribution $p(x|\theta)$ (e.g., $\theta \in \mathbb{R}^k$). Goal: Find the parameter $\hat{\theta}$ that maximizes the probability of the observed data, i.e.,

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, p(x_1, \ldots, x_N|\theta) = \underset{\theta}{\operatorname{argmax}} \prod_{n=1}^{N} p(x_i|\theta)$$

Here, $p(x_1, \ldots, x_N|\theta)$ is called the likelihood function, which is a function of θ. $\hat{\theta}$ is called the maximum likelihood estimate (MLE).

General Case

Maximum Likelihood Estimation

Given some dataset $x_1, \ldots, x_N$ drawn independently from some distribution $p(x|\theta)$ (e.g., $\theta \in \mathbb{R}^k$). Goal: Find the parameter $\hat{\theta}$ that maximizes the probability of the observed data, i.e.,

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, p(x_1, \ldots, x_N|\theta) = \underset{\theta}{\operatorname{argmax}} \prod_{n=1}^{N} p(x_i|\theta)$$

Here, $p(x_1, \ldots, x_N|\theta)$ is called the likelihood function, which is a function of θ. $\hat{\theta}$ is called the maximum likelihood estimate (MLE). We can also consider

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \sum_{n=1}^{N} \log p(x_i|\theta),$$

where $l(\theta) = \sum_{n=1}^{N} \log p(x_i|\theta)$ is called the log likelihood of the data.

Even More General Case

Maximum Likelihood Estimation

The data $x_1, \ldots, x_N$ can follow any parameterised distribution we like as long as we can specify the joint probability density function of the data, $p(x_1, \ldots, x_N|\theta)$ (e.g., $\theta \in \mathbb{R}^k$).

Goal: We still want to find the parameter $\hat{\theta}$ that maximizes the probability of the observed data, i.e.,

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, p(x_1, \ldots, x_N|\theta)$$

Warning: MLE might not provide a unique solution and MLE solution might not exist!

Modelling data probability distribution

Where does $p(x_1, \ldots, x_N|\theta)$ come from?

We have obtained the dataset $x_1, \ldots, x_N$ and we assume that we can model the probability distribution of the data with the model probability distribution $p(x_1, \ldots, x_N|\theta)$.
That is, it is up to us to choose an appropriate model!

Example

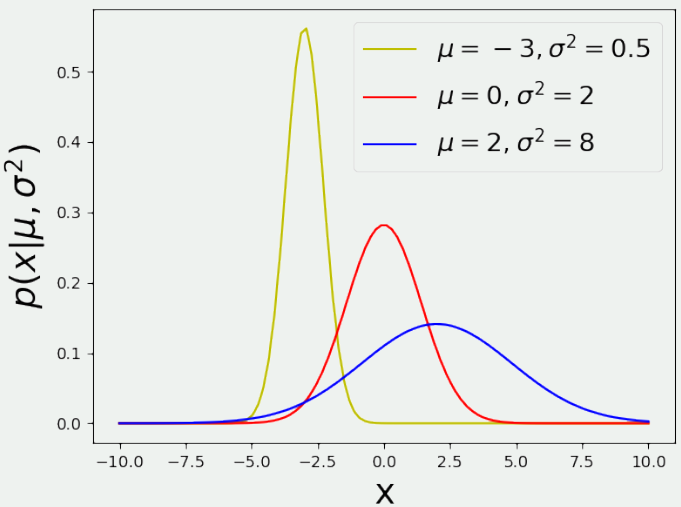
Normal Distribution

The probability density function (pdf) of a normal distribution (a.k.a Gaussian distribution) is given by

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

For a random variable $X \in \mathbb{R}$ with such a pdf, we write $X \sim \mathcal{N}(\mu,\sigma^2)$.

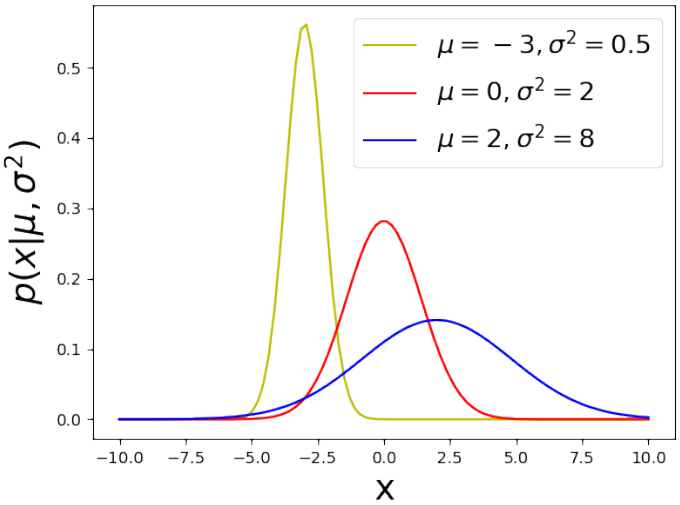
- The parameter μ is called the mean. The pdf achieves is maximum at μ.
- The parameter σ^2 is called the variance. It controls the width of the density function. The value σ is called the standard deviation.



Left as homework

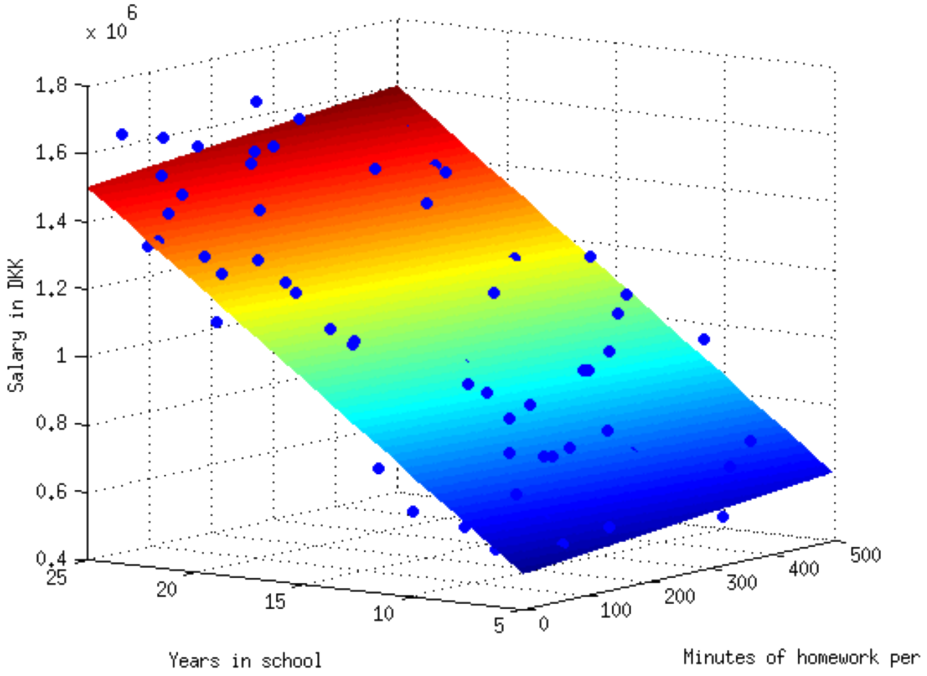
MLE for Gaussian (Exercise 2.8 from the course textbook)

Assume that a dataset $x_1, \ldots, x_N$ consisting of N points was sampled from a Gaussian distribution, i.e., $X_i \sim \mathcal{N}(\mu, \sigma^2)$ for some unknown $-\infty < \mu < \infty$ and unknown $0 < \sigma^2 < \infty$. Also, assume that the X_i are independent and identically distributed (iid). Find the maximum likelihood estimate of the Gaussian mean μ and variance σ^2 and show that the critical point obtained is, at least, a local maximum.



Outline

Multivariate Linear Regression



General Form

- Given: Pairs of the form $(\mathbf{x}_1, t_1), \ldots, (\mathbf{x}_N, t_N) \in \mathbb{R}^D \times \mathbb{R}$.
- Goal: Linear model $f(\boldsymbol{x}; \boldsymbol{w}) = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_D x_D = \boldsymbol{w}^T \boldsymbol{x}$, $\boldsymbol{w}, \boldsymbol{x} \in \mathbb{R}^{D+1}$
  - Don't get confused :-): $\mathbf{x}_n \in \mathbb{R}^D$ and $x_i \in \mathbb{R}$
  - And to make the notation more confusing: $\boldsymbol{x} \in \mathbb{R}^{D+1}$ since $\boldsymbol{x} = (1, x_1, \ldots, x_D)^T$

Polynomial Models

- Let's focus again on $D = 1$, i.e., on input data of the form $x_n \in \mathbb{R}$.
- A non-linear polynomial model can be written on matrix form ...

$$\boldsymbol{X} = \begin{bmatrix} x_1^0 & x_1^1 & x_1^2 & \dots & x_1^K \\ x_2^0 & x_2^1 & x_2^2 & \dots & x_2^K \\ \vdots & & & & \\ x_N^0 & x_N^1 & x_N^2 & \dots & x_N^K \end{bmatrix} \quad \text{and} \quad \boldsymbol{t} = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{bmatrix}$$

- Our model function can then be written as $f(x; \boldsymbol{w}) = \sum_{k=0}^{K} w_k x^k$

Arbitrary "Basis Functions"

- We can basically resort to arbitrary functions ...

$$\boldsymbol{X} = \begin{bmatrix} h_0(x_1) & h_1(x_1) & h_2(x_1) & \dots & h_K(x_1) \\ h_0(x_2) & h_1(x_2) & h_2(x_2) & \dots & h_K(x_2) \\ \vdots & & & & \\ h_0(x_N) & h_1(x_N) & h_2(x_N) & \dots & h_K(x_N) \end{bmatrix} \quad \text{and} \quad \boldsymbol{t} = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{bmatrix}$$

- Our model function can then be written as $f(x; \boldsymbol{w}) = \sum_{k=0}^{K} w_k h_k(x)$
- Notice, this model is non-linear in x, but linear in $\boldsymbol{w}$.

General Case

Also, given more input variables ($D > 1$), we can simply

- transform each input variable/column ...
- combine different input variables (e.g., difference between columns) ...
- combine and transform input variables ...
- ...

Outline

Regression: Statistical Point of View

- Generative models: Measured data = underlying process + random noise

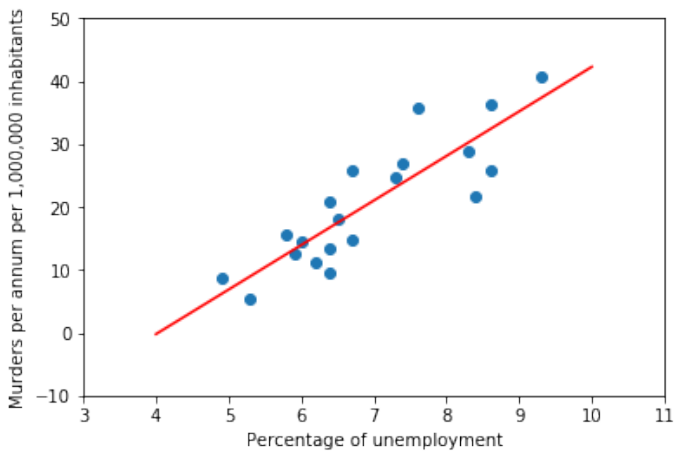


Figure: Observed murder rates are considered a combination of a generative model and random noise.

Regression: Statistical Point of View

- Generative models: Measured data = underlying process + random noise
- Idea: Modelling the noise leads to better models fitting the data?

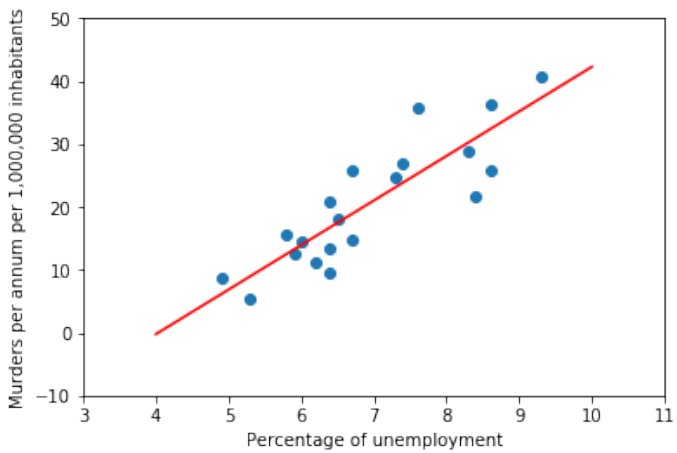


Figure: Observed murder rates are considered a combination of a generative model and random noise.

Regression: Statistical Point of View

- Generative models: Measured data = underlying process + random noise
- Idea: Modelling the noise leads to better models fitting the data?
- Let's consider a new linear regression model of the form

$$t_n = f(\boldsymbol{x}_n, \boldsymbol{w}) + \varepsilon_n,$$

where $\varepsilon_n \sim \mathcal{N}(0, \sigma^2)$ for some σ^2. Note: This makes t_n a random variable!

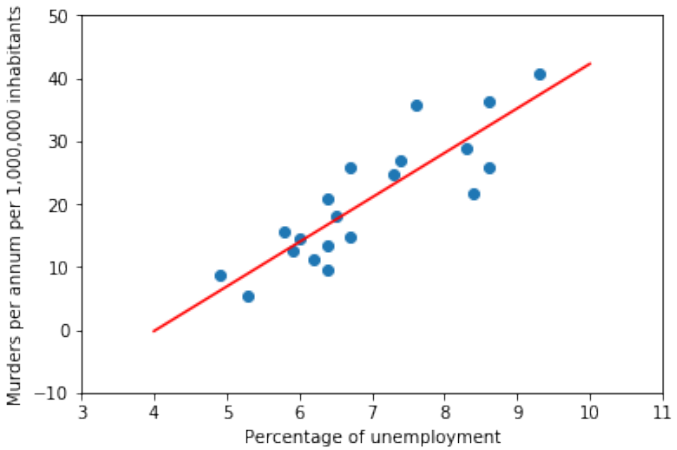


Figure: Observed murder rates are considered a combination of a generative model and random noise.

Random Variable t_n

Gaussians and Shifted Means (Rogers & Girolami)

Adding a constant to a Gaussian random variable is equivalent to another Gaussian random variable with the mean shifted by the same constant:

- $y = a + z$
- $p(z) = \mathcal{N}(m, s)$
- $p(y) = \mathcal{N}(m + a, s)$

Random Variable t_n

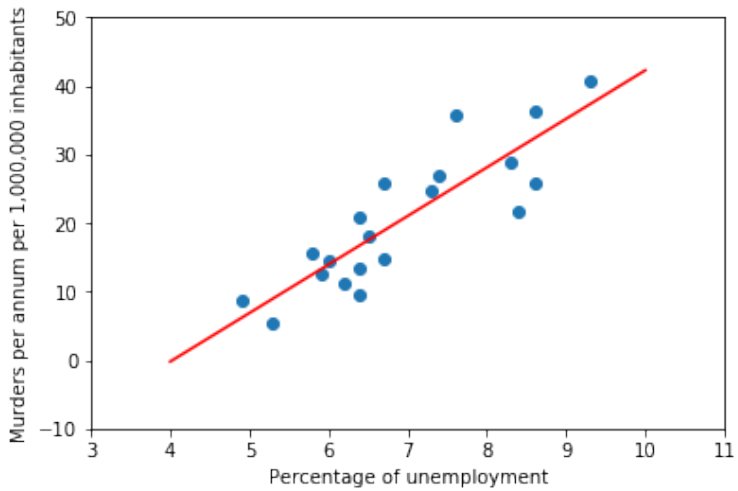
Gaussians and Shifted Means (Rogers & Girolami)

Adding a constant to a Gaussian random variable is equivalent to another Gaussian random variable with the mean shifted by the same constant:

- $y = a + z$
- $p(z) = \mathcal{N}(m, s)$
- $p(y) = \mathcal{N}(m + a, s)$

Thus, our new random variable t_n has the density function (assuming a linear model)

$$p(t_n|\mathbf{x}_n, \mathbf{w}, \sigma^2) = \mathcal{N}(\mathbf{w}^T\mathbf{x}_n, \sigma^2)$$



Random Variable t_n

Gaussians and Shifted Means (Rogers & Girolami)

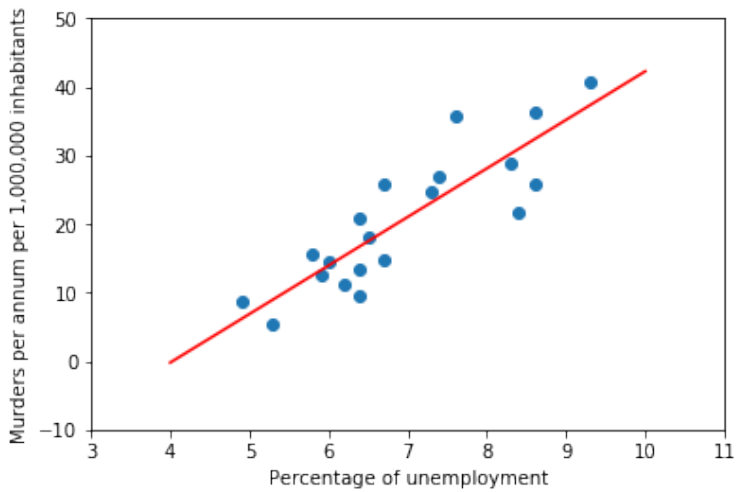
Adding a constant to a Gaussian random variable is equivalent to another Gaussian random variable with the mean shifted by the same constant:

- $y = a + z$
- $p(z) = \mathcal{N}(m, s)$
- $p(y) = \mathcal{N}(m + a, s)$

Thus, our new random variable t_n has the density function (assuming a linear model)

$$p(t_n|\mathbf{x}_n, \mathbf{w}, \sigma^2) = \mathcal{N}(\mathbf{w}^T\mathbf{x}_n, \sigma^2)$$

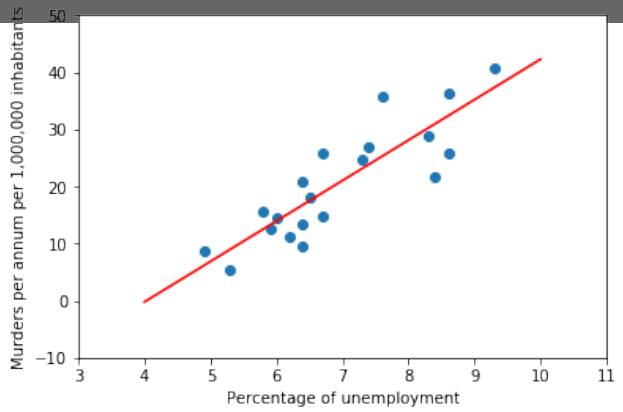
Note that the density $p(t_n|\mathbf{x}_n, \mathbf{w}, \sigma^2)$ of t_n depends on the particular values of $\mathbf{x}_n$ and $\mathbf{w}$ (which determine the mean) and σ^2 (variance); hence, we "condition" on them.



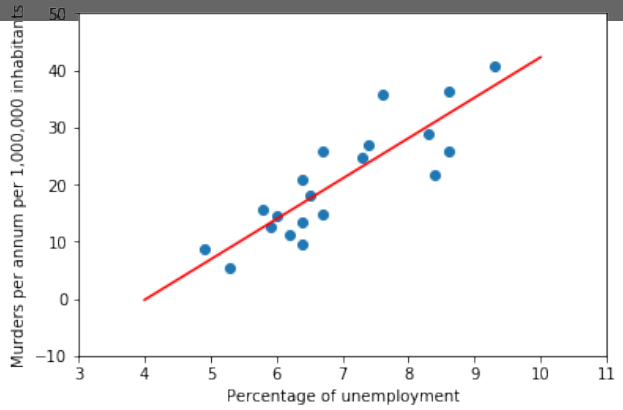
The likelihood of an observed t_n

Our new random variable t_n has the density function

$$p(t_n|\mathbf{x}_n, \mathbf{w}, \sigma^2) = \mathcal{N}(\mathbf{w}^T\mathbf{x}_n, \sigma^2)$$



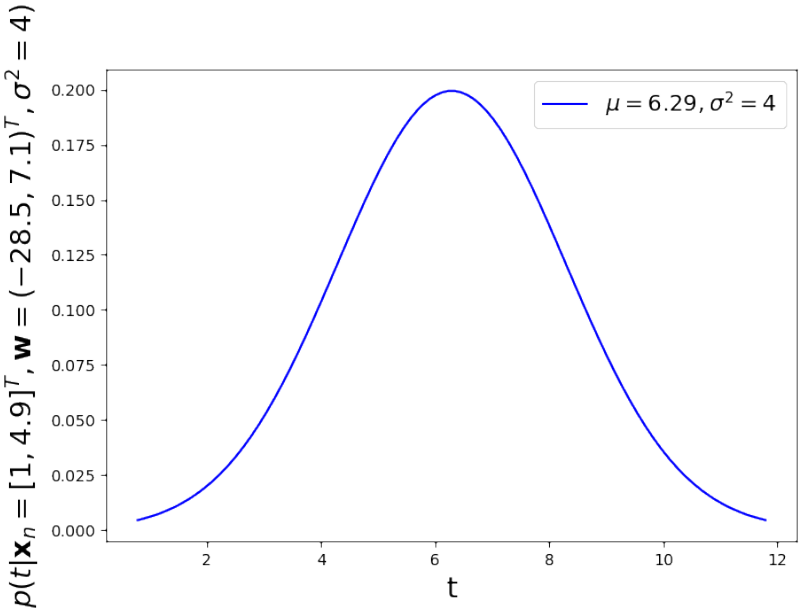
The likelihood of an observed t_n



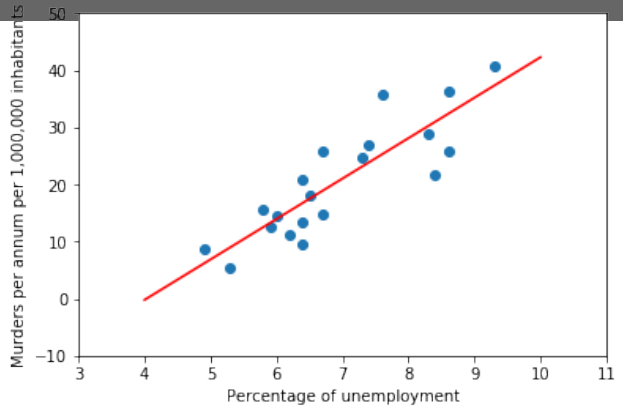
Our new random variable t_n has the density function

$$p(t_n|\mathbf{x}_n, \mathbf{w}, \sigma^2) = \mathcal{N}(\mathbf{w}^T\mathbf{x}_n, \sigma^2)$$

Let's consider a specific model: $\mathbf{w} = (-28.5, 7.1)^T$. Further, let's consider a particular data point $(t_n, x_n) = (8.7, 4.9)$ and a fixed $\sigma^2 = 4$.



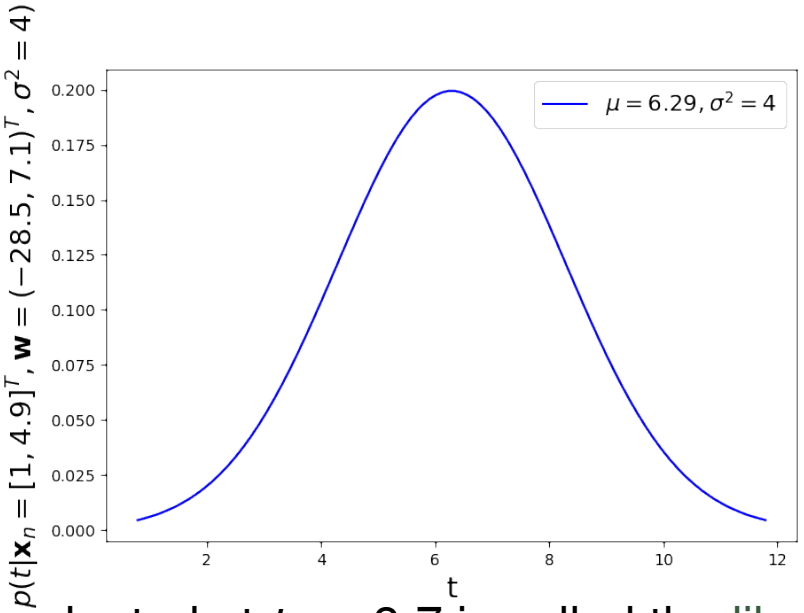
The likelihood of an observed t_n



Our new random variable t_n has the density function

$$p(t_n|\mathbf{x}_n, \mathbf{w}, \sigma^2) = \mathcal{N}(\mathbf{w}^T\mathbf{x}_n, \sigma^2)$$

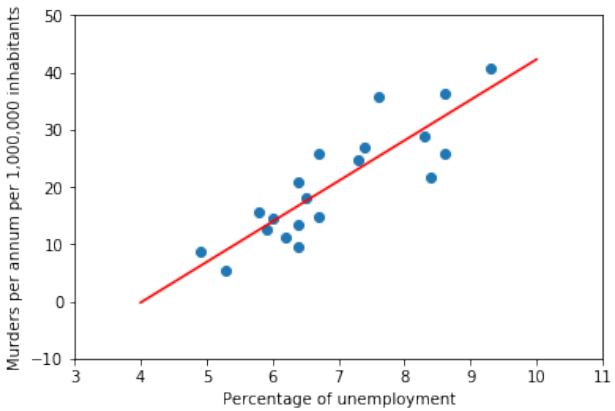
Let's consider a specific model: $\mathbf{w} = (-28.5, 7.1)^T$. Further, let's consider a particular data point $(t_n, x_n) = (8.7, 4.9)$ and a fixed $\sigma^2 = 4$.



The above density evaluated at $t_n = 8.7$ is called the likelihood of the n-th data point. We cannot change $t_n = 8.7$ (our data!), but we can change $\mathbf{w}$ and σ^2 to make the likelihood as high as possible!

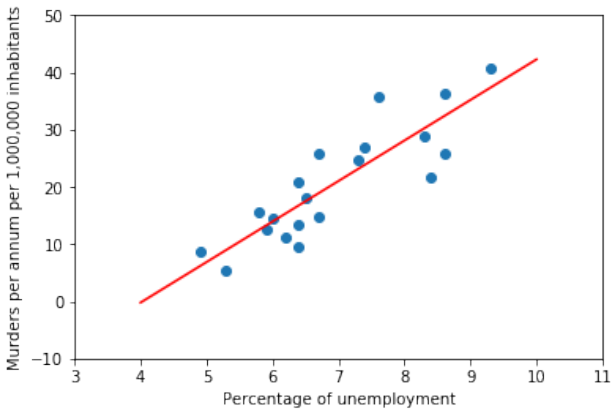
The likelihood of the observed dataset

- We need to simultaneously maximize the likelihood of all the data points
- $p(t_1, \ldots, t_N | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{w}, \sigma^2) = ?$



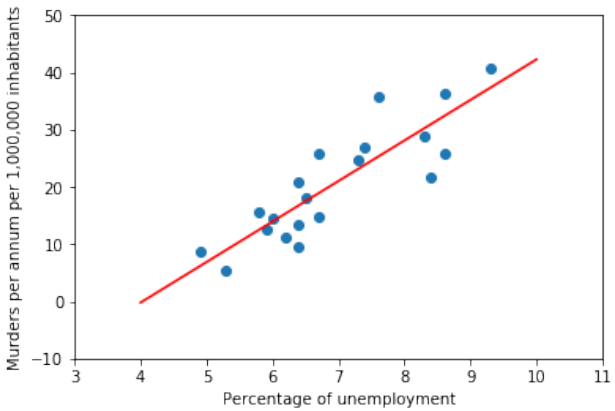
The likelihood of the observed dataset

- We need to simultaneously maximize the likelihood of all the data points
- $p(t_1, \ldots, t_N | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{w}, \sigma^2) = ?$
- A joint density – we make another assumption: Our noise ε_n is i.i.d. (independently and identically distributed)



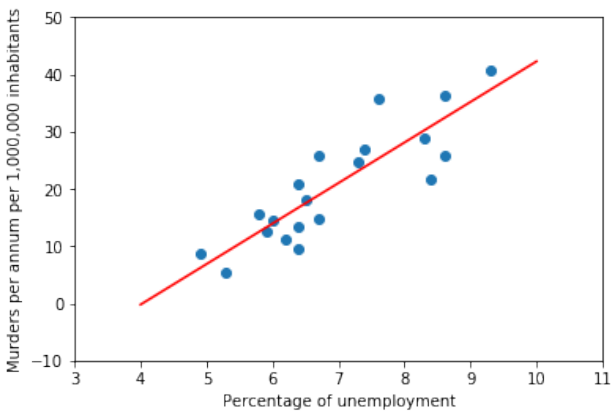
The likelihood of the observed dataset

- We need to simultaneously maximize the likelihood of all the data points
- $p(t_1, \ldots, t_N | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{w}, \sigma^2) = ?$
- A joint density – we make another assumption: Our noise ε_n is i.i.d. (independently and identically distributed)
- $L(\boldsymbol{w}, \sigma^2) = p(t_1, \ldots, t_N | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{w}, \sigma^2) = \prod_{n=1}^{N} p(t_n | \boldsymbol{x}_n, \boldsymbol{w}, \sigma^2)$



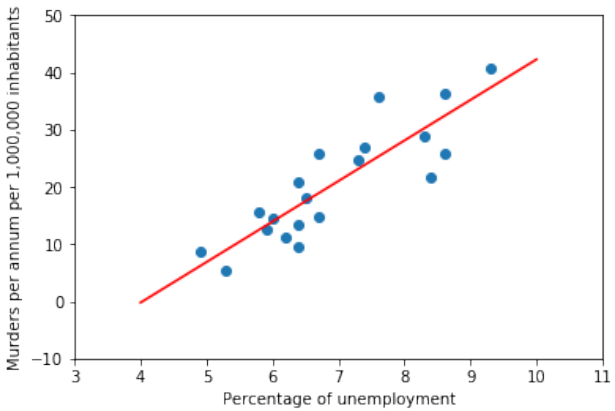
The likelihood of the observed dataset

- We need to simultaneously maximize the likelihood of all the data points
- $p(t_1, \ldots, t_N | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{w}, \sigma^2) = ?$
- A joint density – we make another assumption: Our noise ε_n is i.i.d. (independently and identically distributed)
- $L(\boldsymbol{w}, \sigma^2) = p(t_1, \ldots, t_N | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{w}, \sigma^2) = \prod_{n=1}^{N} p(t_n | \boldsymbol{x}_n, \boldsymbol{w}, \sigma^2)$
- Why did I not assume that the t_n were i.i.d?



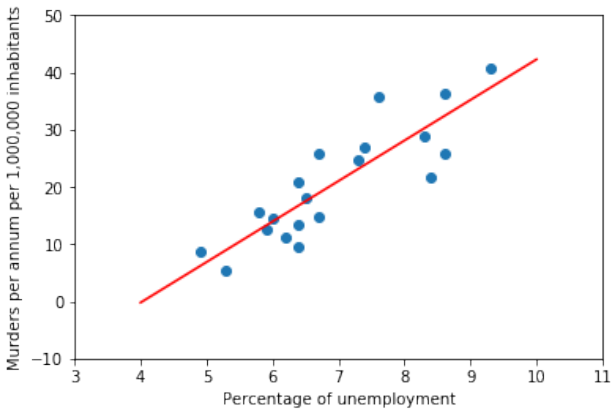
The likelihood of the observed dataset

- We need to simultaneously maximize the likelihood of all the data points
- $p(t_1, \ldots, t_N | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{w}, \sigma^2) = ?$
- A joint density – we make another assumption: Our noise ε_n is i.i.d. (independently and identically distributed)
- $L(\boldsymbol{w}, \sigma^2) = p(t_1, \ldots, t_N | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{w}, \sigma^2) = \prod_{n=1}^{N} p(t_n | \boldsymbol{x}_n, \boldsymbol{w}, \sigma^2)$
- Why did I not assume that the t_n were i.i.d?
- They are not! There is a deterministic part of the model, which means the t_n are very much dependent ...



The likelihood of the observed dataset

- We need to simultaneously maximize the likelihood of all the data points
- $p(t_1, \ldots, t_N | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{w}, \sigma^2) = ?$
- A joint density – we make another assumption: Our noise ε_n is i.i.d. (independently and identically distributed)
- $L(\boldsymbol{w}, \sigma^2) = p(t_1, \ldots, t_N | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{w}, \sigma^2) = \prod_{n=1}^{N} p(t_n | \boldsymbol{x}_n, \boldsymbol{w}, \sigma^2)$
- Why did I not assume that the t_n were i.i.d?
- They are not! There is a deterministic part of the model, which means the t_n are very much dependent ...
- New task: $\operatorname{argmax}_{(\boldsymbol{w}, \sigma^2)} L(\boldsymbol{w}, \sigma^2) = \prod_{n=1}^{N} p(t_n | \boldsymbol{x}_n, \boldsymbol{w}, \sigma^2)$



Maximum Likelihood: Derivations I

- $\text{argmax}_{(\boldsymbol{w},\sigma^2)} L(\boldsymbol{w},\sigma^2) = \prod_{n=1}^{N} p(t_n|\boldsymbol{x}_n,\boldsymbol{w},\sigma^2) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{w}^T\mathbf{x}_n,\sigma^2)$

Maximum Likelihood: Derivations I

- $\text{argmax}_{(\boldsymbol{w},\sigma^2)} L(\boldsymbol{w},\sigma^2) = \prod_{n=1}^{N} p(t_n|\boldsymbol{x}_n,\boldsymbol{w},\sigma^2) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{w}^T\mathbf{x}_n,\sigma^2)$
- Let's fix σ^2 first. Further, let us consider the log-likelihood

$$\log L(\mathbf{w}) = \sum_{n=1}^{N} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(t_n-\mathbf{w}^T\mathbf{x}_n)^2\right\}\right)$$

Maximum Likelihood: Derivations I

- $\text{argmax}_{(\boldsymbol{w},\sigma^2)} L(\boldsymbol{w},\sigma^2) = \prod_{n=1}^{N} p(t_n|\boldsymbol{x}_n,\boldsymbol{w},\sigma^2) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{w}^T\mathbf{x}_n,\sigma^2)$
- Let's fix σ^2 first. Further, let us consider the log-likelihood

$$\begin{aligned}
\log L(\mathbf{w}) &= \sum_{n=1}^{N} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(t_n-\mathbf{w}^T\mathbf{x}_n)^2\right\}\right) \\
&= \sum_{n=1}^{N}\left(-\frac{1}{2}\log(2\pi)-\log\sigma-\frac{1}{2\sigma^2}(t_n-\mathbf{w}^T\mathbf{x}_n)^2\right)
\end{aligned}$$

Maximum Likelihood: Derivations I

- $\text{argmax}_{(\boldsymbol{w},\sigma^2)} L(\boldsymbol{w},\sigma^2) = \prod_{n=1}^{N} p(t_n|\boldsymbol{x}_n,\boldsymbol{w},\sigma^2) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{w}^T\mathbf{x}_n,\sigma^2)$
- Let's fix σ^2 first. Further, let us consider the log-likelihood

$$\begin{aligned}
\log L(\mathbf{w}) &= \sum_{n=1}^{N} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(t_n-\mathbf{w}^T\mathbf{x}_n)^2\right\}\right) \\
&= \sum_{n=1}^{N}\left(-\frac{1}{2}\log(2\pi)-\log\sigma-\frac{1}{2\sigma^2}(t_n-\mathbf{w}^T\mathbf{x}_n)^2\right) \\
&= -\frac{N}{2}\log(2\pi)-N\log\sigma-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(t_n-\mathbf{w}^T\mathbf{x}_n)^2
\end{aligned}$$

Maximum Likelihood: Derivations I

- $\text{argmax}_{(\boldsymbol{w},\sigma^2)} L(\boldsymbol{w},\sigma^2) = \prod_{n=1}^{N} p(t_n|\boldsymbol{x}_n,\boldsymbol{w},\sigma^2) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{w}^T\mathbf{x}_n,\sigma^2)$
- Let's fix σ^2 first. Further, let us consider the log-likelihood

$$\begin{aligned}
\log L(\mathbf{w}) &= \sum_{n=1}^{N} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(t_n-\mathbf{w}^T\mathbf{x}_n)^2\right\}\right) \\
&= \sum_{n=1}^{N}\left(-\frac{1}{2}\log(2\pi)-\log\sigma-\frac{1}{2\sigma^2}(t_n-\mathbf{w}^T\mathbf{x}_n)^2\right) \\
&= -\frac{N}{2}\log(2\pi)-N\log\sigma-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(t_n-\mathbf{w}^T\mathbf{x}_n)^2
\end{aligned}$$

- Next, let's consider the gradient with respect to the parameters $\boldsymbol{w}$

$$\nabla \log L(\mathbf{w}) = \frac{1}{2\sigma^2}\sum_{n=1}^{N} 2(t_n-\mathbf{w}^T\mathbf{x}_n)\mathbf{x}_n = \frac{1}{\sigma^2}\sum_{n=1}^{N}\mathbf{x}_n t_n - \mathbf{x}_n\mathbf{x}_n^T\mathbf{w}$$

Maximum Likelihood: Derivations II

- $\nabla \log L(\mathbf{w}) = \frac{1}{2\sigma^2} \sum_{n=1}^{N} 2(t_n - \mathbf{w}^T \mathbf{x}_n)\mathbf{x}_n = \frac{1}{\sigma^2} \sum_{n=1}^{N} \mathbf{x}_n t_n - \mathbf{x}_n \mathbf{x}_n^T \mathbf{w}$

Maximum Likelihood: Derivations II

- $\nabla \log L(\mathbf{w}) = \frac{1}{2\sigma^2} \sum_{n=1}^{N} 2(t_n - \mathbf{w}^T \mathbf{x}_n)\mathbf{x}_n = \frac{1}{\sigma^2} \sum_{n=1}^{N} \mathbf{x}_n t_n - \mathbf{x}_n \mathbf{x}_n^T \mathbf{w}$
- One can rewrite this as (see textbook, Exercise 1.5)

$$\nabla \log L(\mathbf{w}) = \frac{1}{\sigma^2} \left(\mathbf{X}^T \mathbf{t} - \mathbf{X}^T \mathbf{X} \mathbf{w} \right)$$

Maximum Likelihood: Derivations II

- $\nabla \log L(\mathbf{w}) = \frac{1}{2\sigma^2} \sum_{n=1}^{N} 2(t_n - \mathbf{w}^T \mathbf{x}_n)\mathbf{x}_n = \frac{1}{\sigma^2} \sum_{n=1}^{N} \mathbf{x}_n t_n - \mathbf{x}_n \mathbf{x}_n^T \mathbf{w}$
- One can rewrite this as (see textbook, Exercise 1.5)

$$\nabla \log L(\mathbf{w}) = \frac{1}{\sigma^2} \left(\mathbf{X}^T \mathbf{t} - \mathbf{X}^T \mathbf{X} \mathbf{w} \right)$$

- Setting the gradient to $\mathbf{0}$ yields

$$\begin{aligned} \frac{1}{\sigma^2} \left(\mathbf{X}^T \mathbf{t} - \mathbf{X}^T \mathbf{X} \mathbf{w} \right) &= \mathbf{0} \\ \mathbf{X}^T \mathbf{t} &= \mathbf{X}^T \mathbf{X} \mathbf{w} \\ \left(\mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{t} &= \mathbf{w} \end{aligned}$$

Maximum Likelihood: Derivations III

- The Hessian $\mathbf{H} = \nabla^2 \log L(\mathbf{w}) = -\frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{X}$ is negative semidefinite. Hence, $\hat{w}$ is a global maximum.
- Note that $\hat{w}$ is optimal for any choice of σ^2!
- Hence, one can now search for an optimal $\widehat{\sigma^2}$ ...
- Textbook: One can show that $\widehat{\sigma^2} = \frac{1}{N}(\mathbf{t}^T\mathbf{t} - \mathbf{t}^T\mathbf{X}\mathbf{w})$ corresponds to a maximum ...

Maximum Likelihood: Derivations III

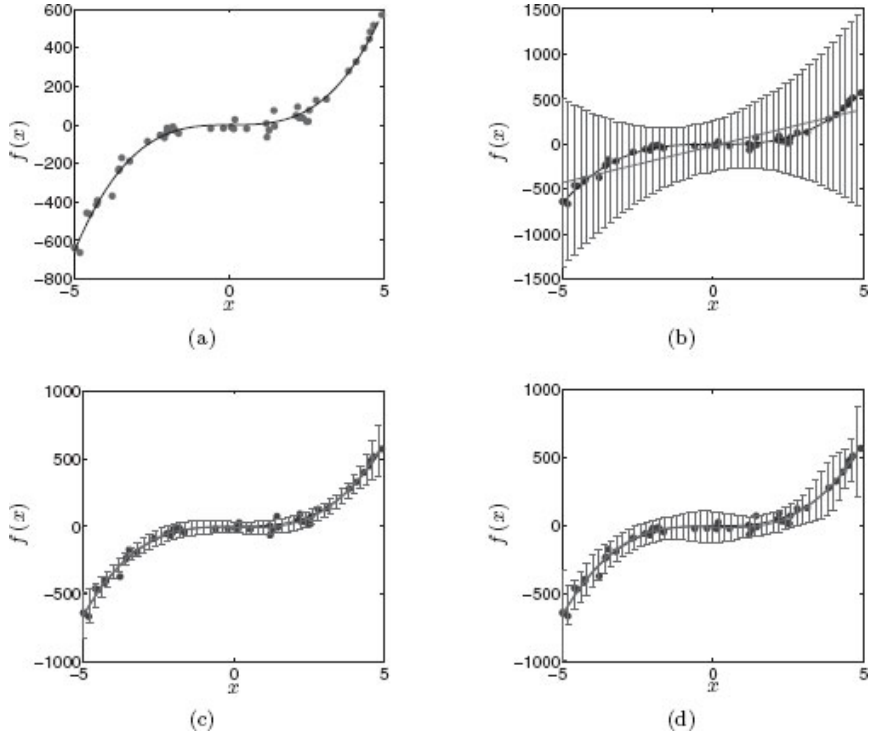
- The Hessian $\mathbf{H} = \nabla^2 \log L(\mathbf{w}) = -\frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{X}$ is negative semidefinite. Hence, $\hat{w}$ is a global maximum.
- Note that $\hat{w}$ is optimal for any choice of σ^2!
- Hence, one can now search for an optimal $\widehat{\sigma^2}$ ...
- Textbook: One can show that $\widehat{\sigma^2} = \frac{1}{N}(\mathbf{t}^T\mathbf{t} - \mathbf{t}^T\mathbf{X}\mathbf{w})$ corresponds to a maximum ...

Maximum Likelihood Estimate

$$\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t}$$

$$\widehat{\sigma^2} = \frac{1}{N}(\mathbf{t}^T\mathbf{t} - \mathbf{t}^T\mathbf{X}\hat{\mathbf{w}})$$

Predictive Variance



Prediction Variability (Rogers & Girolami, Section 2.11)

For a new set of attributes $\mathbf{x}_{new}$, one can compute a prediction t_{new} as well as the associated variance $\sigma^2_{new} = \mathbf{var}(t_{new})$:

1. $t_{new} = \mathbf{x}^T_{new}\hat{\mathbf{w}}$
2. $\sigma^2_{new} = \widehat{\sigma^2}\mathbf{x}^T_{new}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_{new}$

to compute the estimate and the associated predictive variance.

Maximum Likelihood & Linear Regression

Maximum Likelihood Estimate

$$\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t}$$

$$\widehat{\sigma^2} = \frac{1}{N}(\mathbf{t}^T\mathbf{t} - \mathbf{t}^T\mathbf{X}\hat{\mathbf{w}})$$

- Probabilistic viewpoint of regression (t_n is now a random variable).
- The MLE of linear regression under an i.i.d. additive Gaussian noise model is equivalent to the linear least squares solution from the first week.
- Potential problems? Potential disadvantages?

Maximum Likelihood & Linear Regression

Maximum Likelihood Estimate

$$\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t}$$

$$\widehat{\sigma^2} = \frac{1}{N}(\mathbf{t}^T\mathbf{t} - \mathbf{t}^T\mathbf{X}\hat{\mathbf{w}})$$

- Probabilistic viewpoint of regression (t_n is now a random variable).
- The MLE of linear regression under an i.i.d. additive Gaussian noise model is equivalent to the linear least squares solution from the first week.
- Potential problems? Potential disadvantages?
  - Overfitting (as before!)

Maximum Likelihood & Linear Regression

Maximum Likelihood Estimate

$$\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t}$$

$$\widehat{\sigma^2} = \frac{1}{N}(\mathbf{t}^T\mathbf{t} - \mathbf{t}^T\mathbf{X}\hat{\mathbf{w}})$$

- Probabilistic viewpoint of regression (t_n is now a random variable).
- The MLE of linear regression under an i.i.d. additive Gaussian noise model is equivalent to the linear least squares solution from the first week.
- Potential problems? Potential disadvantages?
  - Overfitting (as before!)
  - MLE is a point estimate. No measure of uncertainty w.r.t. $\mathbf{w}$ and σ.

Maximum Likelihood & Linear Regression

Maximum Likelihood Estimate

$$\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t}$$

$$\widehat{\sigma^2} = \frac{1}{N}(\mathbf{t}^T\mathbf{t} - \mathbf{t}^T\mathbf{X}\hat{\mathbf{w}})$$

- Probabilistic viewpoint of regression (t_n is now a random variable).
- The MLE of linear regression under an i.i.d. additive Gaussian noise model is equivalent to the linear least squares solution from the first week.
- Potential problems? Potential disadvantages?
  - Overfitting (as before!)
  - MLE is a point estimate. No measure of uncertainty w.r.t. $\mathbf{w}$ and σ.
  - MLE might not exist . . .

Maximum Likelihood & Linear Regression

Maximum Likelihood Estimate

$$\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t}$$

$$\widehat{\sigma^2} = \frac{1}{N}(\mathbf{t}^T\mathbf{t} - \mathbf{t}^T\mathbf{X}\hat{\mathbf{w}})$$

- Probabilistic viewpoint of regression (t_n is now a random variable).
- The MLE of linear regression under an i.i.d. additive Gaussian noise model is equivalent to the linear least squares solution from the first week.
- Potential problems? Potential disadvantages?
  - Overfitting (as before!)
  - MLE is a point estimate. No measure of uncertainty w.r.t. $\mathbf{w}$ and σ.
  - MLE might not exist ...
  - MLE might not be unique ...

Maximum Likelihood & Linear Regression

Maximum Likelihood Estimate

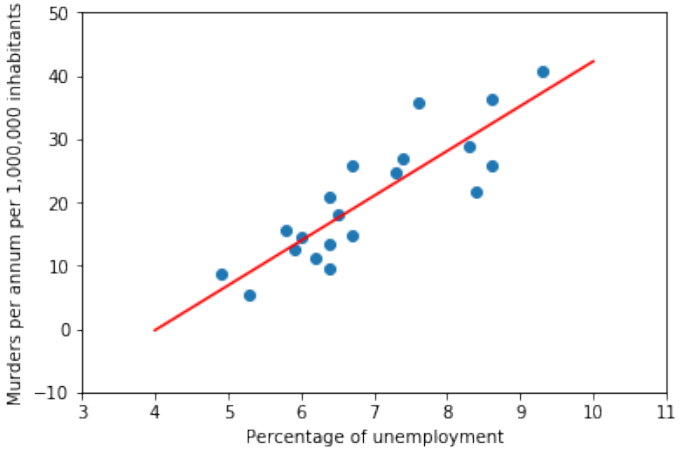
$$\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t}$$

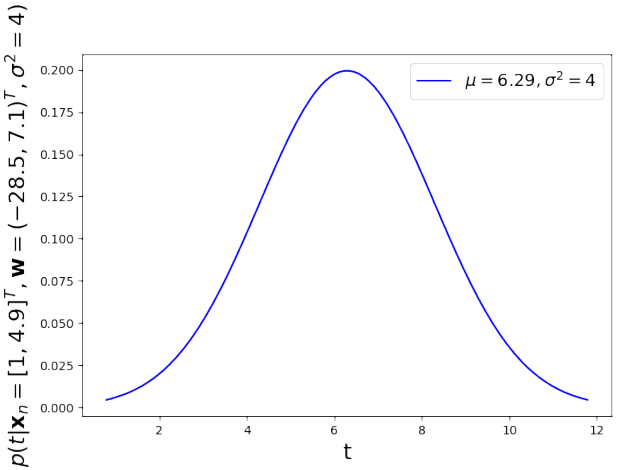
$$\widehat{\sigma^2} = \frac{1}{N}(\mathbf{t}^T\mathbf{t} - \mathbf{t}^T\mathbf{X}\hat{\mathbf{w}})$$

- Probabilistic viewpoint of regression (t_n is now a random variable).
- The MLE of linear regression under an i.i.d. additive Gaussian noise model is equivalent to the linear least squares solution from the first week.
- Potential problems? Potential disadvantages?
  - Overfitting (as before!)
  - MLE is a point estimate. No measure of uncertainty w.r.t. $\mathbf{w}$ and σ.
  - MLE might not exist ...
  - MLE might not be unique ...
- Next: Bayesian perspective!

Outline

Summary & Outlook





Today

- Overview & Next Steps
- Maximum likelihood estimation in general
- Recap: Multivariate linear regression models
- Maximum likelihood estimation & regression

Outlook

- Learn about a Bayesian approach to regression, resulting in a regression model with a model of uncertainty (on Thursday)
- Learn how to sample from distributions and do inference (after the Christmas break)