Linear Modeling - Maximum Likelihood

TF4063

Fadjar Fathurrahman

The material in this note is based on [1].

[1]

1 Random variables and probability

Random variables

Random variables can be described as variables whose values depend on outcomes of random events. There are two kinds of random variables: discrete and continuous random variables. Discrete random variables are used for random events for which we can systematically list (or count) all possible outcomes. Continuous random variables are used when we can not list all possible outcomes. It is a common convention to use upper-case letters to describe random variables and lower-case ones for possible values that the random variable can take.

A discrete random variable could be used to describe, for examples:

- a coin toss: possible outcomes are head or tail.
- rolling of a die: possible outcomes are 1,2,3,4,5 or 6.

A continous random variable could be used to describe, for examples:

- outcome of a 100m race
- height of a human
- mass of a pebble

The collection of possible outcomes is known as the sample space.

Probability and distributions

Let Y be a random variable that represents the toss of a coin. If the coin lands heads, $Y = 1$ and if tails, $Y = 0$. To model this event (the coin toss), we need to be able to quantify how likely either outcome is. For discrete random variables, we do this by defining *probabilities* of different outcomes.

Two important rules governing probabilities:

- Probabilities must be greater than or equal to 0 and less than of equal to 1.
$$0 \leq P(Y = y) \leq 1 \tag{1}$$
- The sum of the probabilities of each possible individual outcome must be equal to 1.
$$\sum_y P(Y = y) = 1 \tag{2}$$

For a coin we have:

$$P(Y = 1) + P(Y = 0) = 1 \tag{3}$$

For a fair coin we have $P(Y = 1) = P(Y = 0) = 0.5$.
We will sometimes use the following shorthand

$$P(Y = y) = P(y) \tag{4}$$

The set of all possible outcomes (all of the ys) and their probabilities, $P(y)$, is known as probability distribution. It tell use how the total probability is distributed over all possible outcomes.

Adding probabilities

Let Y be a random variable for modeling outcome of rolling of a fair die. For exmple we want to know the probability of the result being lower than 4. The outcomes that are lower than 4 are 1, 2, and 3, which means that we need to be able to calculate the probability that the die lands 1 or 2 or 3. We can calculate this by using the additive law of probability:

$$P(Y < 4) = P(Y = 1) + P(Y = 2) + P(Y = 3)$$

Conditional probabilities

Often one event will affect the outcome of another. We can use conditional probabilities to express the probability that Y takes a particular value given that X has taken a particular value. This can be expressed as
$$P(Y = y|X = x) \tag{5}$$

or using shorthand notation $P(y|x)$. This notation also can be read as the probability that Y has the outcome of y given that X has the outcome x.

Joint probabilities

Given two (or more) random variables, we want to know the probability that they each take a particular value. For example the probability that Y has the outcome y and X has the outcome x is written as

$$P(Y = y, X = y) \tag{6}$$

or in shorthand notation $p(x, y)$. How we deal with these joint distributions depends on whether or not the random variables are dependent. If the events are independent we have

$$P(y_1, y_2, \ldots, y_j) = \prod_{j=1}^{J} P(y_j) \tag{7}$$

It the events are dependent, we cannot decompose the joint probability as products of individual probabilities. However, if we can create conditional distributions, we can decompose the joint probability using the following definitions

$$P(Y = y, X = x) = P(Y = y | X = x) P(X = x) \tag{8}$$

or

$$P(Y = y, X = x) = P(X = x | Y = y) P(Y = y) \tag{9}$$

Marginalization

Given joint probability $P(Y = y, X = x)$ we can obtain $P(Y = y)$ by marginalizing out X from the joint distribution. This can be done by summing the joint probabilities over all possible values of X:

$$P(Y = y) = \sum_x P(Y = y, X = x) \tag{10}$$

More generally we have

$$P(Y_j = y_j) = P(y_j) = \sum_{y_1, \ldots, y_{j-1}, y_{j+1}, \ldots, y_J} P(y_1, y_2, \ldots, y_J) \tag{11}$$

Bayes' rule

Bayes' rule can be used for reversing conditioning of probability.

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)} \tag{12}$$

Expectations

An expectation tells us what value we would expect some function $f(X)$ of a random variable X to take and is defined (for discrete random variables) as

$$\mathbf{E}_{P(x)}\{f(X)\} = \sum_x f(x)\,P(x) \tag{13}$$

A common expectation that we will encounter is the mean, which is the expectation of $f(X) = X$:

$$\mathbf{E}_{P(x)}\{X\} = \sum_x x\,P(x) \tag{14}$$

For a fair die, for example, $P(x) = 1/6$, we have

$$\mathbf{E}_{P(x)}\{X\} = \sum_x x\frac{1}{6} = \frac{1}{6} + \frac{2}{6} + \ldots + \frac{6}{6} = \frac{21}{6} = 3.5 \tag{15}$$

Another example, for a die, the expected value of $f(X) = X^2$ is

$$\mathbf{E}_{P(x)}\{X^2\} = \sum_x x^2\frac{1}{6} = \frac{1}{6} + \frac{4}{6} + \cdots + \frac{36}{6} = \frac{91}{6} \tag{16}$$

The expected value of a function X is not in general the function evaluated at the expected value of X. Mathematically

$$\mathbf{E}_{P(x)}\{f(X)\} \neq f(\mathbf{E}_{P(x)}\{X\}), \text{ (generally)} \tag{17}$$

One situation where the two are equal is when the function is just a constant multiplied by X. For example, $f(X) = aX$

$$\begin{aligned}\mathbf{E}_{P(x)}\{f(X)\} &= \sum_x axP(x)\\ &= a\sum_x xP(x)\\ &= a\mathbf{E}_{P(x)}\{X\}\\ &= f\left(\mathbf{E}_{P(x)}\{X\}\right)\end{aligned}$$

Another important case is when the function is simply a constant, $f(X) = a$. In this case, the expectation disappears due to the fact that the distribution has to sum to 1 over all possible outcomes

$$\begin{aligned}\mathbf{E}_{P(x)}\{f(X)\} &= \sum_x aP(x) \\ &= a\sum_x P(x) \\ &= a\end{aligned}$$

Expectation of a sum of different functions is equal to a sum of the individual expectations:

$$\begin{aligned}\mathbf{E}_{P(x)}\{f(X)+g(X)\} &= \sum_x (f(x)+g(x))P(x) \\ &= \sum_x f(x)P(x) + \sum_x g(x)P(x) \\ &= \mathbf{E}_{P(x)}\{f(X)\} + \mathbf{E}_{P(x)}\{g(X)\}\end{aligned}$$

Besides mean, another common expectation value that we will encounter is the variance. Variance is a measure of how variable the random variable is and is defined as the expected square deviation from the mean.

$$\mathrm{var}\{X\} = \mathbf{E}_{P(x)}\left\{\left(X - \mathbf{E}_{P(x)}\{X\}\right)^2\right\} \tag{18}$$

Expand the terms in bracket:

$$\begin{aligned}\mathrm{var}\{X\} &= \mathbf{E}_{P(x)}\left\{X^2 - 2X\mathbf{E}_{P(x)}\{X\} + \mathbf{E}_{P(x)}\{X\}^2\right\} \\ &= \mathbf{E}_{P(x)}\{X^2\} - 2\mathbf{E}_{P(x)}\{X\}\mathbf{E}_{P(x)}\{X\} + \mathbf{E}_{P(x)}\{X\}^2\end{aligned}$$

Collecting together the $\mathbf{E}_{P(x)}\{X\}^2$ terms gives

$$\mathrm{var}\{X\} = \mathbf{E}_{P(x)}\{X^2\} - \mathbf{E}_{P(x)}\{X\}^2 \tag{19}$$

Vector random variables

Probability distributions over vectors is nothing more than a shorthand way of defining large joint distributions. For example the values that could be taken on by random variables $X_1, X_2, \ldots, X_N$ can be expressed as the vector $\mathbf{x} = [x_1, x_2, \ldots, x_N]^\mathsf{T}$. Using this shorthand:

$$p(\mathbf{x}) = p(x_1, x_2, \ldots, x_N) = P(X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N) \tag{20}$$

Even though $\mathbf{x}$ is a vector, $p(\mathbf{x})$ is a scalar quantity.

Expectations are computed for vector random variables in the same way.

$$\mathbf{E}_{P(x)}\{f(\mathbf{x})\} = \sum_{\mathbf{x}} f(\mathbf{x})P(\mathbf{x}) \tag{21}$$

where the sum is over all possible values of the vector **x**.

The mean vector is defined as

$$\mathbf{E}_{P(\mathbf{x})} = \sum_{\mathbf{x}} \mathbf{x}\, P(\mathbf{x}) \tag{22}$$

When dealing with vectors, the concept of variance is generalized to a covariance matrix. This is defined as

$$\text{cov}\{\mathbf{x}\} = \mathbf{E}_{P(\mathbf{x})}\left\{\left(\mathbf{x} - \mathbf{E}_{P(\mathbf{x})}\mathbf{x}\right)\left(\mathbf{x} - \mathbf{E}_{P(\mathbf{x})}\mathbf{x}\right)^{\mathsf{T}}\right\} \tag{23}$$

If **x** is a vector of length D, then cov$\{\mathbf{x}\}$ is a $D \times D$ matrix. The diagonal elements correspond to the variance of the individual elements of **x**.

The off-diagonal elements tell us to what extent different elements of **x** co-vary, that is, how dependent they are on one another.

A high positive value between, say, elements x_d and x_e suggest that if x_d increases, so does x_e. A high negative value suggests that they are related but move in opposite directions. A value of or close to zero suggest that there is no relationship between them.

The covariance for vector random variables also can be written as

$$\begin{aligned}\text{cov}\{\mathbf{x}\} &= \mathbf{E}_{P(\mathbf{x})}\left\{\left(\mathbf{x} - \mathbf{E}_{P(\mathbf{x})}\{\mathbf{x}\}\right)\left(\mathbf{x} - \mathbf{E}_{P(\mathbf{x})}\{\mathbf{x}\}\right)^{\mathsf{T}}\right\} \\ &= \mathbf{E}_{P(\mathbf{x})}\left\{\mathbf{x}\mathbf{x}^{\mathsf{T}} - 2\mathbf{x}\mathbf{E}_{P(\mathbf{x})}\{\mathbf{x}\}^{\mathsf{T}} - \mathbf{E}_{P(\mathbf{x})}\{\mathbf{x}\}\mathbf{E}_{P(\mathbf{x})}\{\mathbf{x}\}^{\mathsf{T}}\right\}\end{aligned}$$

We finally obtain

$$\text{cov}\{\mathbf{x}\} = \mathbf{E}_{P(\mathbf{x})}\left\{\mathbf{x}\mathbf{x}^{\mathsf{T}}\right\} - \mathbf{E}_{P(\mathbf{x})}\{\mathbf{x}\}\mathbf{E}_{P(\mathbf{x})}\{\mathbf{x}^{\mathsf{T}}\} \tag{24}$$

2 Popular discrete distributions

Bernoulli distribution

For a random variable X that can take two values, 0 or 1 (a binary random variable), where the probability that it takes the value 1 is defined as q, the Bernoulli distribution is

$$P(X = x) = q^x(1-q)^{1-x} \tag{25}$$

Binomial distribution

The binomial distribution extends the Bernoulli distribution to define the probability of observing a certain number of heads in a total of N tosses. More generally, we might think of events that have two outcomes (success or failure). If we have N such events, the binomial random variable Y can take the values from 0 (no success) to N (N success). The probability of observing a particular number of successes is given by:

$$P(Y = y) = P(y) = \begin{pmatrix} N \\ y \end{pmatrix} q^y (1 - q)^{N-y} \tag{26}$$

where:

$$\begin{pmatrix} N \\ y \end{pmatrix} = \frac{N!}{y!(N - y)!} \tag{27}$$

is a mathematical shorthand for the number of ways in which y distinct objects can be chosen from a set of N objects.

The Bernoulli distribution is a special case of the binomial distribution when $N = 1$.

Multinomial distribution

Multinomial distribution is a generalization of binomial distribution to vector random variables.

$$P(Y = \mathbf{y}) = P(\mathbf{y}) = \frac{N!}{\prod_j y_j!} \prod_j q_j^{y_j} \tag{28}$$

q_j are the parameters of the multinomial distribution

$$\sum_j q_j = 1$$

3 *Continuous random variables - density functions*

When working with continuous random variables, we need a continuous analogue to the probability distribution. This is provided by a probability density function (pdf) or simply *density*, also denoted $p(x)$. To compute the probability that X lies in a particular range, we compute the definite integral of $p(x)$ with respect x over this range

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} p(x) \, \mathrm{d}x$$

If a random variable X only takes values in the range $x_1 \leq X \leq x_2$ we have

$$\int_{x_1}^{x_2} p(x)\,\mathrm{d}x = 1, \text{ where } x_1 \leq X \leq x_2 \tag{29}$$

We also have

$$p(x) \geq 0 \tag{30}$$

Note that there is no upper bound on the value of pdf because it is not a probability and can be higher than 1 for a particular value of x.

We also can define joint pdf over several continuous random variables. For example $p(x, y)$ the joint density of two random variables X and Y and $p(\mathbf{w})$ is the density of a vector $\mathbf{w}$ which could be thought of as the joint density of $p(w_0, w_1, \ldots)$ random variables representing each element in the vector. Although we cannot compute $P(X = x, Y = y)$ we can compute

$$P(x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2) = \int_{x=x_1}^{x=x_2} \int_{y=y_1}^{y=y_2} p(x, y)\,\mathrm{d}x\,\mathrm{d}y \tag{31}$$

The same applies for conditional distributions, although the conditioning is done on exact value (as this event is assume to have happened or known). For example we can compute

$$P(x_1 \leq X \leq x_2 | Y = y) = \int_{x=x_1}^{x=x_2} p(x|Y = y)\,\mathrm{d}x$$

For marginalization, we can use integration instead of summation. For example the pdf $p(y)$ can be computed from $p(y, x)$ as follows:

$$p(y) = \int_{x=x_1}^{x=x_2} p(y, x)\,\mathrm{d}x$$

where $x_1 \leq X \leq x_2$ described the sample space of X.

Expectations with respect to continous random variables are performed by integrating over the range of values that the random variable can take:

$$\mathbf{E}_{p(x)}\{f(x)\} = \int f(x)p(x)\,\mathrm{d}x \tag{32}$$

For many cases, this integration cannot be performed analytically. In this case, we can approximate it using simple summation

$$\mathbf{E}_{p(x)}\{f(x)\} \approx \frac{1}{S}\sum_{s=1}^{S} f(x_s)P(x_s) \tag{33}$$

4 Popular continuous density function

Uniform density function

$$p(y) = \begin{cases} r & \text{for } a \leq y \leq b \\ 0 & \text{otherwise} \end{cases} \tag{34}$$

Given a and b, the value of r can be calculated from the condition

$$P(a \leq Y \leq b) = \int_{y=a}^{y=b} p(y)\,\mathrm{d}y = 1$$

So we have

$$r = \frac{1}{b-a}$$

Beta density function

Beta density function can be used for continuous random variables that are restricted to between 0 and 1. The beta density function is defined as

$$p(r) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} r^{\alpha-1}(1-r)^{\beta-1} \tag{35}$$

where α and β are positive parameters that control the shape of the density function. $\Gamma(z)$ is known as the gamma function.

Gaussian or normal density function

Gaussian random variables are used in many continuous applications. It is useful because it can be manipulated easily in several situations. Gaussian density function is defined over a sample space that includes all real numbers. It is usually defined as

$$p(y|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\} \tag{36}$$

where the conditioning is done over two variables: the mean (μ) and variance σ^2. Gaussian density function also can be written in shorthand notation as

$$p(y|\mu,\sigma^2) = \mathcal{N}(\mu,\sigma^2) \tag{37}$$

In Python Numpy we can use `np.random.randn` to draw sample from standard normal distribution, with $\mu = 1$ and $\sigma = 1$. For random samples from $\mathcal{N}(\mu,\sigma^2)$ we can use

```
sigma * np.random.randn(...) + mu
```

Multivariate Gaussian

The Gaussian density function also can be generalized for continous vectors.

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma}|^{1/2}} \exp\left\{ -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \right\} \tag{38}$$

In Python, we can use `numpy.random.multivariate_normal` to draw samples from multivariate normal distribution.

5 *Likelihood*

We will take into account error in our data by considering the following model.

$$t_n = \mathbf{w}^{\mathsf{T}}\mathbf{x}_n + \epsilon_n \tag{39}$$

ϵ_n is a continuous random variable. We do not just have one random variable, but one for each observed data. We assume that these values are independent:

$$p(\epsilon_1, \epsilon_2, \ldots, \epsilon_N) = \prod_{n=1}^{N} p(\epsilon_n) \tag{40}$$

We additionally assume the form of $p(\epsilon_n)$ is that of Gaussian distribution with zero mean and variance σ^2.
The model now consists of two components:

1. Deterministic component: $\mathbf{w}^{\mathsf{T}}\mathbf{x}_n$, sometimes referred to as a trend or drift.

2. Random component: ϵ_n, sometimes referred to as noise.

Our model is of the following form:

$$t_n = f(\mathbf{x}_n; \mathbf{w}) + \epsilon_n,\ \epsilon_n \sim \mathcal{N}(0, \sigma^2) \tag{41}$$

We need to find the optimal value of $\mathbf{w}$, $\widehat{\mathbf{w}}$, such that this model describes our data as best as possible. The loss measured the difference between the observed values of t and those predicted by the model. The effect of adding a random variable to the model is that the output of the model, t, is now itself a random variable. In other words,

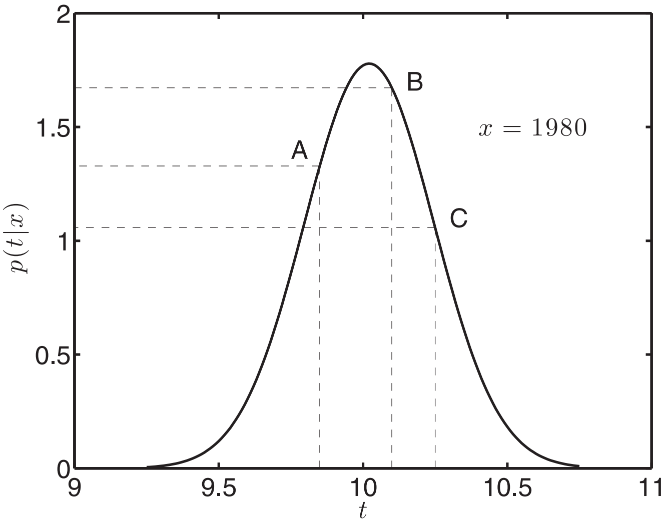
there is no single value of t_n for a particular x_n. As such, we cannot use the loss as a means of to find $\mathbf{w}$ and σ_2
Adding a constant $\mathbf{w}^\mathsf{T}\mathbf{x}_n$ to a Gaussian random variable is equivalent to another Gaussian random variable with the mean shifted by the same constant:

$$
\begin{aligned}
y &= a + z \\
p(z) &= \mathcal{N}(\mu, \sigma^2) \\
p(y) &= \mathcal{N}(\mu + a, \sigma^2)
\end{aligned}
$$

Therefore, the random variable t_n has the density function:

$$p(t_n|\mathbf{x}, \mathbf{w}, \sigma^2) = \mathcal{N}(\mathbf{w}^\mathsf{T}\mathbf{x}_n, \sigma^2) \tag{42}$$

Note the conditioning on the left hand side. The density of t_n depends on particular value of $\mathbf{x}_n$ and $\mathbf{w}$ and also σ^2 (the variance). As an example, we will plot the likelihood function from one data point of `olympic100m` dataset. We will chose the data for year 1980 and evaluate the likelihood using $\mathbf{w}$ obtained from minimizing loss function and assuming that $\sigma^2 = 0.05$.



Recall that, for a continuous random variable, t, $p(t)$ cannot be interpreted as a probability. The height of the curve at a particular value of t can be interpreted as how likely it is that we would observe that particular t for $x = 1980$. The most likely winning time in 1980 would be 10.02 seconds (for a Gaussian, the most likely (highest) point corresponds to the mean). Also shown on the plot, are three example times – A, B and C. Of these, B is the most likely and C the least likely.

The actual winning time in the 1980 Olympics is C (10.25 seconds). The density $p(t_n|\mathbf{x}_n, \mathbf{w}, \sigma^2)$ evaluated at t n = 10.25 is an important quantity, known as the likelihood of the nth data point. We cannot change $t_n = 10.25$ (this is our data) but we can change $\mathbf{w}$ and σ^2 to try and move the density so as to make it as high as possible at $t = 10.25$. The idea of finding parameters that maximize the likelihood in this way is a key concept in Machine Learning.

Dataset likelihood

In general we are not interested in the likelihood of single data point but that of all of the data. If we have N data points we have the following joint conditional density

$$p(t_1, \ldots, t_N|\mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{w}, \sigma^2)$$

This is a joint density over all of the responses in our dataset. We will write this compactly as $p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2)$. Evaluating this density at the observed data points gives a single likelihood value for the whole dataset, which we can optimise by varying $\mathbf{w}$ and σ^2.

The assumption that the noise at each data point is independent, i.e.

$$p(\epsilon_1, \ldots, \epsilon_N) = \prod_n p(\epsilon_n)$$

enables us to factorize this density into something more manageable. In particular, this joint conditional density can be factorized into N separate terms, one for each data object:

$$L = p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} p(t_n|\mathbf{x}_n, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{w}^T\mathbf{x}_n, \sigma^2) \qquad (43)$$

For analytical reasons, we will maximise the natural logarithm of the likelihood. We will follow the Machine Learning convention of using $\log(y)$ to denote the natural logarithm of y, often denoted elsewhere as $\ln(y)$). We can do this because the estimated arguments $\widehat{\mathbf{w}}$ and $\widehat{\sigma^2}$ that maximize the log-likelihood will also maximize the likelihood.

We will start by writing the log-likelihood as

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

For our choice of model $f(\mathbf{x}_n; \mathbf{w}) = \mathbf{w}^\mathsf{T} \mathbf{x}_n$, we have

$$\log L = -\frac{N}{2} \log 2\pi - N \log \sigma - \frac{1}{2\sigma^2} \sum_{n=1}^{N} \left(t_n - \mathbf{w}^\mathsf{T} \mathbf{x}_n\right)^2 \tag{44}$$

First derivative w.r.t $\mathbf{w}$:

$$\begin{aligned}
\frac{\partial \log L}{\partial w} &= \frac{1}{\sigma^2} \sum_{n=1}^{N} \mathbf{x}_n \left(t_n - \mathbf{x}_n^\mathsf{T} \mathbf{w}\right) \\
&= \frac{1}{\sigma^2} \sum_{n=1}^{N} \mathbf{x}_n t_n - \mathbf{x}_n \mathbf{x}_n^\mathsf{T} \mathbf{w} = \mathbf{0}
\end{aligned}$$

we have used $\mathbf{w}^\mathsf{T} \mathbf{x}_n = \mathbf{x}_n^\mathsf{T} \mathbf{w}$ Using matrix-vector notation

$$\begin{aligned}
\sum_{n=1}^{N} \mathbf{x}_n t_n &= \mathbf{X}^\mathsf{T} \mathbf{t} \\
\sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^\mathsf{T} \mathbf{w} &= \mathbf{X}^\mathsf{T} \mathbf{X} \mathbf{w}
\end{aligned}$$

we can write the derivative as

$$\frac{\partial \log L}{\partial w} = \frac{1}{\sigma^2} \left(\mathbf{X}^\mathsf{T} \mathbf{t} - \mathbf{X}^\mathsf{T} \mathbf{X} \mathbf{w} = \mathbf{0} \right) \tag{45}$$

solving this equation we can find the optimal value $\widehat{\mathbf{w}}$

$$\widehat{\mathbf{w}} = \left(\mathbf{X}^\mathsf{T} \mathbf{X} \right)^{-1} \mathbf{X}^\mathsf{T} \mathbf{t} \tag{46}$$

This is the maximum likelihood solution for $\mathbf{w}$ and this solution is exactly the same as the solution obtained by minimizing the loss function. Minimizing the squared loss is equivalent to the maximum likelihood solution if the noise is assumed to be Gaussian.

To obtain the expression for σ^2, we can use the same procedure, assuming that $\mathbf{w} = \widehat{\mathbf{w}}$:

$$\frac{\partial \log L}{\partial \sigma} = -\frac{N}{\sigma} + \frac{1}{\sigma^3} \sum_{n=1}^{N} (t_n - \mathbf{x}^\mathsf{T} \widehat{\mathbf{w}})^2 = 0 \tag{47}$$

Rearranging the equation we have

$$\widehat{\sigma^2} = \frac{1}{N} \sum_{n=1}^{N} (t_n - \mathbf{x}^\mathsf{T} \widehat{\mathbf{w}})^2 \tag{48}$$

This expression can be simplified to

$$\widehat{\sigma^2} = \frac{1}{N} \left(\mathbf{t}^\mathsf{T}\mathbf{t} - \mathbf{t}^\mathsf{T}\mathbf{X}\widehat{\mathbf{w}} \right) \tag{49}$$

Using `olympic100m` data we can obtain these parameters as

```
w = [36.416455902505334, -0.013330885710962845]
σ2 = 0.05030711047565789
```

Maximum likelihood favors complex models

Plugging the expression for $\widehat{\sigma^2}$ into the log-likelihood expression gives use the value of log-likelihood at the maximum:

$$\log L = -\frac{N}{2}(1 + \log 2\pi) - \frac{N}{2} \log \widehat{\sigma^2} \tag{50}$$

This tells us that the maximum value of L will keep increasing as we decrease $\widehat{\sigma^2}$.

The more complex model is overfitting - we have given the model too much freedom and it is attempting to make sense out of what is essentially noise. We showed how regularization could be used to penalise overcomplex parameter values. The same can be done with probabilistic models through the use of prior distributions on the parameter values.

6 *Effect of noise on parameter estimates*

Uncertainty in estimates

The value of $\widehat{\mathbf{w}}$ is strongly influenced by the particular noise values in the data. It would be useful to know how much uncertainty there was in $\widehat{\mathbf{w}}$. In other words, is this $\widehat{\mathbf{w}}$ is unique in explaining the data well or are there many that could do almost as well?

$$t_n = \mathbf{w}^{\mathsf{T}}\mathbf{x}_n + \epsilon_n \tag{51}$$

where $\mathbf{w}$ represents the true value of the parameters and ϵ_n is a random variable that we have defined to be normally distributed. This assumption means that the generating distribution of likelihood is a product of normal densities:

$$p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2) = \prod_{n=1}^{N} p(t_n|\mathbf{x}_n,\mathbf{w},\sigma^2) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{w}^{\mathsf{T}}\mathbf{x}_n,\sigma^2) \tag{52}$$

It is more convenient to work with multivariate Gaussian

$$p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2) = \mathcal{N}(\mathbf{X}\mathbf{w},\sigma^2\mathbf{I}) \tag{53}$$

Now, $\widehat{\mathbf{w}}$ is an estimate of the true parameter value $\mathbf{w}$. Computing the expectation of $\widehat{\mathbf{w}}$ w.r.t the generating distribution will tell us what we expect $\widehat{\mathbf{w}}$ to be on average, and using $\widehat{\mathbf{w}} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{t}$, we have

$$\begin{aligned}
\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\{\widehat{\mathbf{w}}\} &= \int \widehat{\mathbf{w}} p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)\,\mathrm{d}\mathbf{t} \\
&= (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}} \\
&= (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\{\mathbf{t}\} \\
&= (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{w} \\
&= \mathbf{w}
\end{aligned}$$

where we have used the fact that the expected value of a normally distributed random variable is equal to its mean $\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\{\mathbf{t}\} = \mathbf{X}\mathbf{w}$. This result tells us that the expected value of our approximation $\widehat{\mathbf{w}}$ is the true parameter value. This means that our estimat for $\mathbf{w}$ is unbiased - it is not, on average, too big or too small.

Potential variability in the estimate of $\widehat{\mathbf{w}}$ is encapsulated in its *covariance matrix*. It can be showed that:

$$\mathrm{cov}\{\widehat{\mathbf{w}}\} = \sigma^2(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1} = -\left(\frac{\partial \log L}{\partial \mathbf{w}\partial \mathbf{w}^{\mathsf{T}}}\right)^{-1} \tag{54}$$

The certainty or uncertainty in the parameters as described by $\mathrm{cov}\widehat{(\mathbf{w})}$ is directly linked to the second derivative of the log likelihood.

Covariance matrix provides us with two useful pieces of information. The diagonal elements (the variances of the individual elements in $\widehat{\mathbf{w}}$) tell us how much variability we might expect in the individual parameters. The off-diagonal elements tell us how the parameters covary – if the values are high and positive, it tells us that increasing one will require an increase in the other to maintain a good model. Large negative values tell us the opposite – increasing one will cause a decrease in the other. Values close to zero tell us that the param eters are not dependent on one another.

Comparison with empirical values

[using numerical experiments]
If we use $\widehat{\mathbf{w}}_s$ to describe the parameters obtained from the s-th dataset, the empirical covariance matrix can be computed as

$$\widehat{\operatorname{cov}\{\widehat{\mathbf{w}}\}} = \frac{1}{S}\sum_{s=1}^{S}\left(\widehat{\mathbf{w}}_s - \widehat{\boldsymbol{\mu}}\right)\left(\widehat{\mathbf{w}}_s - \widehat{\boldsymbol{\mu}}\right)^{\mathsf{T}}$$

where

$$\widehat{\boldsymbol{\mu}} = \frac{1}{S}\sum_{s=1}^{S}\widehat{\mathbf{w}}_s$$

7 *Variability in predictions*

Making predictions

To predict t_{new}, we multiply $\mathbf{x}_{\text{new}}$ by the best set of model parameters, $\widehat{\mathbf{w}}$

$$t_{\text{new}} = \mathbf{x}_{\text{new}}^{\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{t} = \mathbf{x}_{\text{new}}^{\mathsf{T}}\widehat{\mathbf{w}} \tag{55}$$

with variance:

$$\sigma_{\text{new}}^2 = \sigma^2\mathbf{x}_{\text{new}}^{\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{x}_{\text{new}} \tag{56}$$

σ^2 is the true variance of the dataset noise. In its place, we can use our estimate, $\widehat{\sigma^2}$.

Estimate of the noise variance

Recall our estimate to variance:

$$\widehat{\sigma^2} = \frac{1}{N}\left(\mathbf{t}^{\mathsf{T}}\mathbf{t} - \mathbf{t}^{\mathsf{T}}\mathbf{X}\widehat{\mathbf{w}}\right) \tag{57}$$

Computing expectations of this expression with respect to $p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)$ we obtain:

$$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\} = \sigma^2\left(1 - \frac{D}{N}\right) \tag{58}$$

where D is the number of columns in $\mathbf{X}$. Assuming that $D < N$ (i.e.

the number of attributes we measure for each data point is smaller than the number of data points), then our estimate of the variance will, on average, be lower than the true variance:

$$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\} < \sigma^2 \tag{59}$$

Unlike $\widehat{\mathbf{w}}$ this estimator is biased.