Stats 102B - Week 5, Lecture 1

Miles Chen, PhD

Department of Statistics

Week 5 Monday

UCLA

Section 1

Review: Least Squares Regression

Multiple Linear Regression

With multiple linear regression, the relationship between our input values of observation n $\mathbf{x}_n$ and the predicted values $\hat{t}_n$ is linear:

$$\hat{t}_n = \mathbf{w}^T \mathbf{x}_n$$

We estimated the parameters in $\mathbf{w}$ by trying to minimize the errors of our model. That is, we selected $\hat{\mathbf{w}}$ so that the mean squared error MSE is minimized:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{N} \sum_{n=1}^{N} (t_n - \hat{t}_n)^2$$

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

Minimizing Loss

The approach that was used to train a linear regression model, is the same appreach we used for training a neural network.

- Define a way for our model to make predictions based on inputs $\mathbf{X}$ and model parameters $\mathbf{w}$
- Define a loss function (generally MSE)
- Estimate model parameters by minimizing the loss function (via calculus or gradient descent)

We will now take a different approach. Rather than minimizing loss between our predictions and the data, we will try to maximize the likelihood of our data in a generative model

Section 2

Maximum Likelihood Estimation

A generative model

In a generative model, we model our data as the result of a process that includes random noise.

For Linear regression, our generative model for the nth observation looks like this:

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

Where ϵ_n represents some random noise.

The random noise is the key difference

This looks very similar to what we had before when we were minimizing the loss function.

$$\hat{t}_n = \mathbf{w}^T \mathbf{x}_n$$

The only difference is that we have included a random noise term ϵ, but the inclusion of this term makes all the difference.

$$\begin{aligned} t_n &= \mathbf{w}^T \mathbf{x}_n \quad &+\epsilon_n \\ t_n &= \hat{t} \quad &+\epsilon_n \end{aligned}$$

In other words:

$$\text{actual value} = \text{model prediction} + \text{random error}$$

Inclusion of random noise makes our model probabilistic

By including random noise, we now have a probabilistic model.

Our observations can now be said to come from some distribution, and we can calculate the likelihood of our data.

Our goal will now be to select model parameters that will **maximize** the likelihood of our data.

The Linear Model

Our linear model is

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

Key assumption: We will model the random error / noise as coming from a normal (Gaussian) distribution with mean 0 and constant variance σ^2.

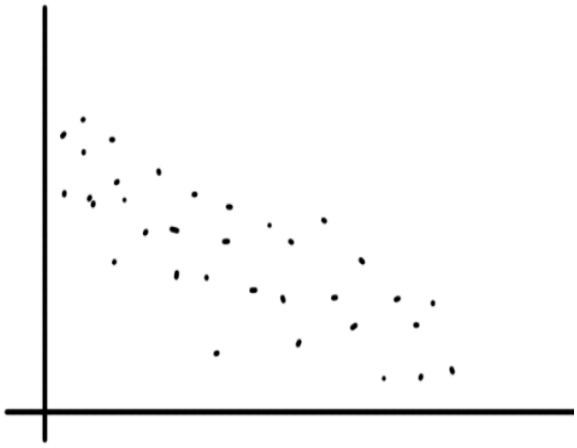
$$\epsilon_n \sim N(0, \sigma^2)$$

When combined with the mean function $f(\mathbf{x}_n; \mathbf{w}_n) = \mathbf{w}^T \mathbf{x}_n$, we have:

$$t_n \sim N(\mathbf{w}^T \mathbf{x}_n, \sigma^2)$$

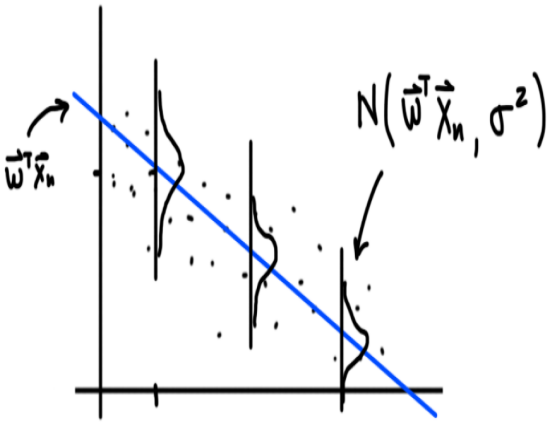
The Linear Model

```
include_graphics("linear_model.png")
```



The Linear Model

```
include_graphics("linear_model_2.png")
```



Maximizing the Likelihood

We now have an probabilistic expression for each observation in our data.

We would like to select our model parameters so that we maximizie the likelihood of observing the data we have.

Section 3

Maximum Likelihood Estimation Review

A quick review

The probability function of observing some value x given some parameter(s) θ is

$$P(x|\theta)$$

Example: Let's say we have a pouch. There are 10 marbles in the pouch. 3 of the marbles are red. The rest are blue. This has a Bernouli distribution. The probability of drawing a red marble is 3/10.

$$P(x = 1|\theta = 0.3) = \theta^x(1 - \theta)^{1-x}$$

$$P(x = 1|\theta = 0.3) = 0.3^1(0.7)^{1-1} = 0.3$$

Probability of independent events is the product of their probabilities.

$$P(x_1, x_2, ...x_N) = \prod_{i=1}^{N} P(x_i|\theta)$$

Example: Let's say we have a pouch. There are 10 marbles in the pouch. 3 of the marbles are red. The rest are blue.

What is the probability of drawing the following sequence: red, red, blue (with replacement)?

$$P(\mathbf{x} = 1, 1, 0|\theta = 0.3) = \prod_{i=1}^{3} P(x_i|\theta = 0.3) = \prod_{i=1}^{3} 0.3^{x_i}(0.7)^{1-x_i} = (0.3)(0.3)(0.7) = 0.063$$

(This is slightly different from the binomial probability function because this question asks for the probability of the exact sequence: red, red, blue.)

Probability vs Likelihood

Probability functions and Likelihood functions both describe probabilities. The distinction is this:

- A probability function is a function of the outcome x. That is, we leave the parameters θ fixed and we can change the outcome.
  - Probability functions must sum to 1 over all possible values of x if discrete, or integrate to 1 (over all possible values of x) if continuous.
- A Likelihood function is a function of the parameter θ. That is, we leave the observed data $\mathbf{x}$ as fixed and we can change the parameter θ.
  - A likelihood function does not need to sum or integrate to 1 over all possible values of θ

The Likelihood function

Let's revisit our pouch with marbles.

This time we have a pouch but don't know exactly how many marbles are red or blue.

We drew three marbles and observed the following sequence: red, red, blue. We will call red 1, and blue 0.

$$L(\theta|\mathbf{x} = 1, 1, 0) = \prod_{i=1}^{3} P(x_i|\theta) = \prod_{i=1}^{3} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^2(1-\theta)$$

Maximum Likelihood Estimation

One strategy to estimate the parameters of a distribution is to maximize the likelihood function associated with our observed data.

In our example, we observed the sequence: red, red, blue.

What is the proportion of red marbles in the pouch that would maximize the likelihood of observing our data?

From the previous slide, we have:

$$L(\theta|\mathbf{x} = 1, 1, 0) = \theta^2(1 - \theta)$$

What value of θ maximizes this?

Maximizing the log-likelihood

Dealing with products can be tricky in calculus, so we will take the logarithm of the likelihood function. The value that maximizes the log-likelihood will be the same value that maximizes the likelihood itself.

$$
\begin{aligned}
\log L &= 2\log\theta + \log(1-\theta) \\
\frac{\partial \log L}{\partial \theta} &= \frac{\partial}{\partial \theta}\left(2\log\theta + \log(1-\theta)\right) \\
0 &= \frac{2}{\theta} + \frac{1}{(1-\theta)}(-1) \\
\frac{1}{(1-\theta)} &= \frac{2}{\theta} \\
\theta &= 2(1-\theta) \\
3\theta &= 2 \\
\theta &= 2/3
\end{aligned}
$$

Back to Linear Regression

With our probabilistic linear regression model, we say our data $\mathbf{t}$ comes from the following distribution:

$$t_n \sim N(\mathbf{w}^T \mathbf{x}_n, \sigma^2)$$

Assuming all of our observations are independent of each other, the likelihood of our data can then be expressed as:

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

The Normal Distribution

The Normal distribution

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

Likelihood function of Linear Regression data

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

We assume all observations of t_n are independent. We assume ϵ is normally distributed $N(0, \sigma^2)$

$$L(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2) = \prod_{n=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-1}{2\sigma^2}(t_n - \mathbf{w}^T\mathbf{x}_n)^2\right)$$

The Likelihood function is a function of $\mathbf{w}$

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

Our likelihod is a function of $\mathbf{w}$.

All other values are treated as constants or fixed values.

- the variance σ^2 is a constant
- the observed values $\mathbf{t}$ are fixed values
- the input values $\mathbf{X}$ are fixed values.

Only the model parameters $\mathbf{w}$ vary in this function.

We want to find the values of $\mathbf{w}$ that will maximize this function.

Maximizing the Likelihood function

To maximize our likelihood function analytically, we will take the derivative, set it equal to 0, and solve for $\mathbf{w}$. Taking the derivative of a product can get messy, so we will maximize the log-likelihood (which turns the product into a sum) instead.

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

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

$$\text{log-likelihood} = \frac{-N}{2}\log(2\pi) - N\log(\sigma) + \frac{-1}{2\sigma^2}\sum_{n=1}^{N}(t_n - \mathbf{w}^T\mathbf{x}_n)^2$$

The solution that maximizes this log-likelihood will be the same value that maximizes the likelihod function. To maximize, we take the derivative, set it to 0, and solve.

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

$$\begin{aligned}
\frac{\partial \log \mathcal{L}}{\partial \mathbf{w}} &= 0 - 0 + \frac{-1}{\sigma^2}\sum_{n=1}^{N}(t_n - \mathbf{w}^T\mathbf{x}_n) \times (-\mathbf{x}_n) \\
&= \frac{1}{\sigma^2}\sum_{n=1}^{N}\underbrace{\mathbf{x}_n}_{2\times 1}(\underbrace{t_n}_{1\times 1} - \underbrace{\mathbf{w}^T\mathbf{x}_n}_{1\times 1}) \\
&= \frac{1}{\sigma^2}\sum_{n=1}^{N}\underbrace{\mathbf{x}_n}_{2\times 1}(\underbrace{t_n}_{1\times 1} - \underbrace{\mathbf{x}_n^T\mathbf{w}}_{1\times 1}) \\
&= \frac{1}{\sigma^2}\sum_{n=1}^{N}(\underbrace{\mathbf{x}_n}_{2\times 1}\underbrace{t_n}_{1\times 1} - \underbrace{\mathbf{x}_n}_{2\times 1}\underbrace{\mathbf{x}_n^T\mathbf{w}}_{1\times 1})
\end{aligned}$$

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

$$\mathbf{x}_n = \begin{bmatrix} 1 \\ x_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_N^T \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix} \qquad \mathbf{t} = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{bmatrix} \qquad \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

$$\sum_{n=1}^{N} \underbrace{\mathbf{x}_n t_n}_{2\times 1} = \underbrace{\mathbf{X}^T}_{2\times N}\underbrace{\mathbf{t}}_{N\times 1} = \underbrace{\mathbf{X}^T\mathbf{t}}_{2\times 1}$$

$$\sum_{n=1}^{N} \underbrace{\mathbf{x}_n}_{2\times 1}\underbrace{\mathbf{x}_n^T}_{1\times 2}\underbrace{\mathbf{w}}_{2\times 1} = \underbrace{\mathbf{X}^T}_{2\times N}\underbrace{\mathbf{X}}_{N\times 2}\underbrace{\mathbf{w}}_{2\times 1} = \underbrace{\mathbf{X}^T\mathbf{X}\mathbf{w}}_{2\times 1}$$

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

Solution

We set the derivative equal to 0 and solve for $\hat{\mathbf{w}}$

$$\frac{\partial \log \mathcal{L}}{\partial \mathbf{w}} = \frac{1}{\sigma^2}(\mathbf{X}^T\mathbf{t} - \mathbf{X}^T\mathbf{X}\mathbf{w}) = 0$$

Thus, the value of $\mathbf{w}$ that maximizes the log-likelihood is

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

This is the exact same solution that minimizes the SSE and MSE, which we derived back in week 1.

Minimizing the squared loss is equivalent to maximizing the likelihood **if the errors are assumed to come from a Gaussian distribution**.

We also note that the variance σ^2 does not appear in the solution at all. It appears only as a constant and does not affect the $\mathbf{w}$ we choose.

Maximizing the log-likelihood function

After we turn the likelihood function into a log-likelihood function, if we look closely, we can see why maximizing the log-likelihood function is equivalent to minimizing the SSE Loss function.

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

$$\text{log-likelihood} = \underbrace{\frac{-N}{2}\log(2\pi) - N\log(\sigma)}_{\text{constants}} + \underbrace{\frac{-1}{2\sigma^2}}_{\text{negative constant}} \underbrace{\sum_{n=1}^{N}(t_n - \mathbf{w}^T\mathbf{x}_n)^2}_{\text{SSE}}$$

In the log-likelihood function, we notice that there are a bunch of constants and the expression for the sum of squared errors is in there. Because the SSE is multiplied by a negative constant, we can see that maximizing the log-likelihood function will be equivalent to minimizing the SSE.