

Statistical Learning Notes

Srikar Katta and Edoardo Airoldi

Contents

Statistical Learning	1
Supervised Learning	2
Linear Regression	2
Unsupervised Learning	10
Latent Space Models with Network Data	10

Statistical Learning

One of the most common applications of the probabilistic models we discussed is for prediction problems. In the corporate world, companies may predict which subscribers will leave the program; in the health care world, doctors may want to know the chances of someone having a specific disease given their diagnostic data; in the social sciences, we may want to predict who the real author behind a speech is. We can solve all of these problems using the probabilistic models we discussed. In fact, the act of developing a predictive model that automatically understands data is known as statistical learning: the machine identifies patterns in the observed data to allow us to predict unknowns. While there are many different flavors of statistical learning, we will approach the field from a probabilistic and statistical perspective. Throughout this section, we will establish likelihood functions for models commonly used in statistical learning, specifically in the areas of supervised and unsupervised learning.

Definition 0.1: Statistical Learning

Statistical learning is the utilization of computational techniques and probabilistic modeling to automatically understand data.

There are a few different subareas of statistical learning. In supervised learning, our goal is to predict some outcome given other observations of that outcome; in other words, the quantity we want to predict is an observed random variable that guides our study of covariation among our data. In unsupervised learning, our goal is to infer the hidden structure of the data; in other words, the quantity we want to predict is a latent variable, and there is no quantity to supervise our study of covariation in our data. And with semi-supervised learning, we have an outcome of interest that is only partially-observed; so some outcomes are latent variables while others are random variables.

Supervised Learning

In the supervised learning framework, we generally have three steps:

1. find the likelihood given the model and data statements
2. estimate the model's parameters/unknowns
3. we evaluate the model's "performance"

Throughout this section, we will discuss steps 1 and 3 and introduce estimation strategies later since parts 1 and 3 go hand in hand: we must be able to recognize which of the models we propose in step one are most representative of the true data generating process. Step 2 on the other hand is its own problem that we can discuss separately. So, after introducing a few supervised learning methods – namely linear regression and state space models – we discuss the topics of training/testing/validation splitting, cross-validation, and evaluation metrics.

Linear Regression

Supervised learning is subset into two types of supervised learning: regression, in which the quantity we want to predict is “continuous,” and classification, in which the quantity we want to predict is “discrete.” Both continuous and discrete are in quotes because continuity in the framework of statistical learning is not the same as continuity in the framework of mathematics. In a regression problem, the actual labels associated with our data may in fact be discrete (i.e., take on a limited number of values), but the actual predictions need not be discrete. For example, if we are predicting age, our observed data may simply be whole numbers describing the number of years a person has lived. However, because we do not mind a real-valued prediction (e.g., 21.5633), the problem is a regression problem. The distinction between discrete and continuous is up to the modeler and requires care and precision that comes only from experience.

Let us define quantities X_1, \dots, X_m —some of which may be random or constant—a single random variable Y , and a random variable ε . In a regression, we assume that the data generating process is some function that combines X_1, \dots, X_m and ε to yield Y . The following few examples are all possible regressions:

1. $Y = \sum_{j=1}^m \beta_j X_j + \varepsilon$ for some constants $\beta_1, \beta_2, \dots, \beta_m$
2. $Y = \beta_0 + \sum_{j=1}^{10} \sum_{i=1}^m \beta_{ij} X_i^j$
3. $\mathbb{P}(Y = 1) = \frac{1}{1 + e^{-\left(\sum_{j=1}^m \beta_j X_j\right)}}$ for some constants β_1, \dots, β_m assuming Y only has values 0 and 1.

One of the most common regression models is the linear regression model. In the linear regression model, we assume there exists a linear relationship between input quantities (often referred to as “covariates” or “features” or “independent variables”) and the output quantity (i.e., the quantity we want to predict, often called the “target feature” or the “dependent variable”). Suppose we have m independent variables; we generally denote these as X_1, \dots, X_m and the dependent variable as Y . And we assume that we have N IID observations. So, for each IID observation i , there is a set of characteristics that describe that observation, denoted as $X_{i1}, X_{i2}, \dots, X_{im}$ and the outcome for that observation Y_i . In the linear regression set up, these

are our only observed quantities. And in our 2x2 table, we also have to consider the variation/constancy of our terms. Even though X_{i1}, \dots, X_{im} are referred to as dependent “variables,” they do not necessarily have to have variation; they can be constants. However, we assume that in a linear regression framework, Y will always be variable.

Recall that we said there exists a linear relationship between Y_i and X_{i1}, \dots, X_{im} . This simply means that a one unit increase in some independent variable X_j will lead to a β_j unit increase in Y ; depending on the situation, we may want to model β_j may be a constant or a latent variable. So, because there is a linear relationship between Y_i and X_{i1}, \dots, X_{im} , we can write down this model as

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} \dots + \beta_m X_{im}.$$

Note that we call β_0 the “constant,” and is akin to the y-intercept in a traditional linear equation (i.e., the b term in $y = mx + b$). When $X_{i1} = X_{i2} = \dots = X_{im} = 0$ (i.e., all our inputs are 0), it is not guaranteed that the y-intercept is also 0. So, we use β_0 as a placeholder for the value of Y when $X_{i1} = X_{i2} = \dots = 0$. Similar to the other β terms, we can model β_0 as a constant or variable, which we decide based on the situation.

The last component of the linear regression model is the “error” term, which is the source of variation in our model. Recall that $\beta_0, \dots, \beta_m, X_{i1}, \dots, X_{im}$ could all be constants, so let us assume this to be the case. Then, since the sum of constants is constant, that would mean that $\beta_0 + \beta_1 X_{i1} + \dots + \beta_m X_{im}$ is also constant. However, we know that Y is a random variable. So, in order to represent the random variation in Y_i , we introduce the term $\varepsilon_i = Y_i - (\beta_0 + \beta_1 X_{i1} + \dots + \beta_m X_{im})$. And generally, rather than imposing a probability distribution on Y_i , we impose a distribution on ε_i . So, there exists a deterministic relationship between Y_i and ε_i ; in other words, if we know Y_i and all other quantities but ε_i , we can compute ε_i . Likewise, if we know ε_i and all other quantities, we can compute Y_i . And now, we can rewrite ε_i to have Y_i on one side of the equation and everything else on the other side to yield

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_m X_{im} + \varepsilon_i.$$

Now, we can represent the linear regression formulation as a model statement and data statement:

for $i = 1, \dots, n$

$X_{i1}, \dots, X_{im} \sim \mathbb{P}(X_1, \dots, X_m)$ if we are modeling X_1, \dots, X_m as variables

$\beta_0, \dots, \beta_m \sim \mathbb{P}(\beta_0, \dots, \beta_m)$ if we are modeling β_0, \dots, β_m as variables $\varepsilon_i \sim \mathbb{P}(\varepsilon)$

$Y_i | \beta_0, \dots, \beta_m, X_{i1}, \dots, X_{im}, \varepsilon_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_m X_{im} + \varepsilon_i.$

And recall that we only observe Y_i and X_{i1}, \dots, X_{im} for all of our N IID observations. So our data table would look as follows:

Using these quantities, we can there set up our 2x2 table and calculate the likelihoods. Let us consider a few examples:

Example 1. Suppose we are given data on a person’s income, represented with Y , and age, denoted by X ,

Table 1: Linear Regression Data Statement

Obs#	Y	X_1	X_2	\dots	X_m
1	y_1	x_{11}	x_{12}	\dots	x_{1m}
2	y_2	x_{21}	x_{22}	\dots	x_{2m}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
N	y_N	x_{N1}	x_{N2}	\dots	x_{Nm}

for N individuals with the following probabilistic model:

$$\begin{aligned}
 &\text{for } i = 1, \dots, n \\
 &\varepsilon_i \sim \text{Normal}(0, \sigma^2) \\
 &Y_i | X_i, \alpha, \beta, \varepsilon_i = \alpha + \beta X_i + \varepsilon_i.
 \end{aligned}$$

Let us find the 2x2 table and the likelihood for this model.

First, notice that we have data on a person's income and their age. So, we only observe the realizations of the random variables Y_1, \dots, Y_N as y_1, \dots, y_n and the constants X_1, \dots, X_N . Even though Y_i 's distribution is not defined, because there exists a deterministic relationship between ε_i and Y_i , we know that Y_i is a random variable. And since there is no deterministic relationship between X_i and another random variable and the distribution of X_i is not explicitly defined, X_i must be a constant. Additionally, the values for σ^2, α, β are all not observed and do not have probability distributions imposed. So they are unknown constants. Then, the 2x2 table would look as follows:

	Observed	Unobserved
Variable	y_1, \dots, y_N	
Constant	X_1, \dots, X_N	α, β, σ^2

And now we can calculate the likelihood. Since we do not have any latent variables, our proper and complete likelihoods are equivalent. So, the likelihood – which represents the probability of our random variables being their realizations given the model's constants – is then

$$\mathbb{P}(Y_1 = y_1, \dots, Y_N = y_N | \alpha, \beta, \sigma^2, X_1, \dots, X_N).$$

Since Y_1, \dots, Y_N are IID random variables, their joint distribution is the product of their marginal distributions. So,

$$\begin{aligned}
 \text{Likelihood} &= \mathbb{P}(Y_1 = y_1, \dots, Y_N = y_N | \alpha, \beta, \sigma^2, X_1, \dots, X_N) \\
 &= \mathbb{P}(Y_1 = y_1 | \alpha, \beta, \sigma^2, X_1, \dots, X_N) \dots \mathbb{P}(Y_N = y_N | \alpha, \beta, \sigma^2, X_1, \dots, X_N) \\
 &= \prod_{i=1}^N \mathbb{P}(Y_i = y_i | \alpha, \beta, \sigma^2, X_1, \dots, X_N).
 \end{aligned}$$

However, here we may run into a problem: we never explicitly defined the distribution of Y_i . Instead,

we can write $\varepsilon_i = y_i - \alpha - \beta X_i \sim \text{Normal}(0, \sigma^2)$. Since Y_i and ε_i are deterministically related, $\mathbb{P}(Y_i = y_i | \alpha, \beta, \sigma^2, X_1, \dots, X_N) = \mathbb{P}(\varepsilon_i = y_i - \alpha - \beta X_i | \alpha, \beta, \sigma^2, X_1, \dots, X_N)$. So,

$$\begin{aligned}
\text{Likelihood} &= \mathbb{P}(Y_1 = y_1, \dots, Y_N = y_N | \alpha, \beta, \sigma^2, X_1, \dots, X_N) \\
&= \mathbb{P}(Y_1 = y_1 | \alpha, \beta, \sigma^2, X_1, \dots, X_N) \dots \mathbb{P}(Y_N = y_N | \alpha, \beta, \sigma^2, X_1, \dots, X_N) \\
&= \prod_{i=1}^N \mathbb{P}(Y_i = y_i | \alpha, \beta, \sigma^2, X_1, \dots, X_N) \\
&= \prod_{i=1}^N \mathbb{P}(\varepsilon_i = y_i - \alpha - \beta X_i | \alpha, \beta, \sigma^2, X_1, \dots, X_N) \\
&= \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2\sigma^2}(y_i - \alpha - \beta X_i)}
\end{aligned}$$

As we discussed earlier, models are simplified representations of some real world process. In Example 1, we assume that age is the only variable that can explain variation in income, but we know many other features may be involved in deciding someone's income, such as education, industry, experience, and location. Let us now consider a linear regression example with multiple variables now.

Example 2. Suppose we are given the following data for person each person i : income (Y_i), age ($X_{i,age}$), education ($X_{i,edu}$), industry ($X_{i,ind}$), years of experience ($X_{i,exp}$), and location ($X_{i,loc}$). For notational simplicity, let $\sum \beta_j X_{i,j}$ represent $\beta_{edu}X_{i,edu} + \beta_{ind}X_{i,ind} + \beta_{exp}X_{i,exp}$. Now, suppose we have the following data generating process:

for $i = 1, \dots, n$

$$X_{i,age} \sim \text{Normal}(\mu, \sigma^2)$$

$$\varepsilon_i \sim \text{Normal}(0, \gamma^2)$$

$$Y_i | X_{i,age}, X_{i,edu}, X_{i,ind}, X_{i,exp}, X_{i,loc}, \beta_0, \beta_{age}, \beta_{edu}, \beta_{ind}, \beta_{exp}, \beta_{loc}, \varepsilon_i = \beta_0 + \beta_{age}X_{i,age} + \sum \beta_j X_{i,j} + \varepsilon_i.$$

Notice, this model statement is very similar to that in Example 1, but with a few differences: obviously, we have more quantities of interest. But additionally, the variable for age now is normally distributed, so it is an observed variable rather than a constant in our 2x2 table. Additionally, because we do not assume that any of the other terms come from a distribution, we can classify all other terms as constants.

In order to find the likelihood, let us write out our 2x2 table:

	Observed	Unobserved
Variable	$y_1, \dots, y_N, X_{1,age}, \dots, X_{N,age}$	
Constant	$X_{1,edu}, \dots, X_{N,edu}, X_{1,ind}, \dots, X_{N,ind}, X_{1,exp}, \dots, X_{N,exp}$	$\beta_0, \beta_{age}, \beta_{edu}, \beta_{ind}, \beta_{exp}, \mu, \sigma^2, \gamma^2$

First, we know the likelihood is the proper probability of observing the given data given the model's constants:

$$\text{Likelihood} = \mathbb{P}(Y_1 = y_1, \dots, Y_N = y_N, X_{1,age} = x_{1,age}, \dots, X_{N,age} = x_{N,age} | \beta_{age}, \sum \beta_j X_{1,j}, \dots, \sum \beta_j X_{N,j}, \mu, \sigma^2, \gamma^2).$$

Now, because each of our N observations are IID, we can rewrite the likelihood of N observations as the

product of the likelihoods of each observation:

$$\begin{aligned}
\text{Likelihood} &= \mathbb{P}(Y_1 = y_1, \dots, Y_N = y_N, X_{1,age} = x_{1,age}, \dots, X_{N,age} = x_{N,age} | \beta_{age}, \sum \beta_j X_{1,j}, \dots, \sum \beta_j X_{N,j}, \mu, \sigma^2, \gamma^2) \\
&= \mathbb{P}(Y_1 = y_1, X_{1,age} = x_{1,age} | \beta_{age}, \sum \beta_j X_{1,j}, \mu, \sigma^2, \gamma^2) \dots \mathbb{P}(Y_N = y_N, X_{N,age} = x_{N,age} | \beta_{age}, \sum \beta_j X_{N,j}, \mu, \sigma^2, \gamma^2) \\
&= \prod_{i=1}^N \mathbb{P}(Y_i = y_i, X_{i,age} = x_{i,age} | \beta_{age}, \sum \beta_j X_{i,j}, \mu, \sigma^2, \gamma^2).
\end{aligned}$$

Now, by Bayes rule, we can rewrite $\mathbb{P}(Y_i = y_i, X_{i,age} = x_{i,age})$ as $\mathbb{P}(Y_i = y_i | X_{i,age} = x_{i,age}) \mathbb{P}(X_{i,age} = x_{i,age})$. So,

$$\begin{aligned}
\text{Likelihood} &= \mathbb{P}(Y_1 = y_1, \dots, Y_N = y_N, X_{1,age} = x_{1,age}, \dots, X_{N,age} = x_{N,age} | \beta_{age}, \sum \beta_j X_{1,j}, \dots, \sum \beta_j X_{N,j}, \mu, \sigma^2, \gamma^2) \\
&= \mathbb{P}(Y_1 = y_1, X_{1,age} = x_{1,age} | \beta_{age}, \sum \beta_j X_{1,j}, \mu, \sigma^2, \gamma^2) \dots \mathbb{P}(Y_N = y_N, X_{N,age} = x_{N,age} | \beta_{age}, \sum \beta_j X_{N,j}, \mu, \sigma^2, \gamma^2) \\
&= \prod_{i=1}^N \mathbb{P}(Y_i = y_i, X_{i,age} = x_{i,age} | \beta_{age}, \sum \beta_j X_{i,j}, \mu, \sigma^2, \gamma^2) \\
&= \prod_{i=1}^N \mathbb{P}(Y_i = y_i | \beta_{age}, X_{i,age} = x_{i,age}, \sum \beta_j X_{i,j}, \gamma^2) \mathbb{P}(X_{i,age} = x_{i,age} | \mu, \sigma^2).
\end{aligned}$$

Since ε_i and Y_i are deterministically related, the Transformation Theorem tells us the distribution of $Y_i \sim \text{Normal}(\beta_0 + \beta_{age}\mu + \sum \beta_j X_{i,j} + \varepsilon_i, \gamma^2)$. So,

$$\begin{aligned}
\text{Likelihood} &= \mathbb{P}(Y_1 = y_1, \dots, Y_N = y_N, X_{1,age} = x_{1,age}, \dots, X_{N,age} = x_{N,age} | \beta_{age}, \sum \beta_j X_{1,j}, \dots, \sum \beta_j X_{N,j}, \mu, \sigma^2, \gamma^2) \\
&= \mathbb{P}(Y_1 = y_1, X_{1,age} = x_{1,age} | \beta_{age}, \sum \beta_j X_{1,j}, \mu, \sigma^2, \gamma^2) \dots \mathbb{P}(Y_N = y_N, X_{N,age} = x_{N,age} | \beta_{age}, \sum \beta_j X_{N,j}, \mu, \sigma^2, \gamma^2) \\
&= \prod_{i=1}^N \mathbb{P}(Y_i = y_i, X_{i,age} = x_{i,age} | \beta_{age}, \sum \beta_j X_{i,j}, \mu, \sigma^2, \gamma^2) \\
&= \prod_{i=1}^N \mathbb{P}(Y_i = y_i | \beta_{age}, X_{i,age} = x_{i,age}, \sum \beta_j X_{i,j}, \gamma^2) \mathbb{P}(X_{i,age} = x_{i,age} | \mu, \sigma^2) \\
&= \prod_{i=1}^N \mathbb{P}(Y_i = \beta_0 + \beta_{age}X_{i,age} + \sum \beta_j X_{i,j} + \varepsilon_i | \beta_{age}, X_{i,age} = x_{i,age}, \sum \beta_j X_{i,j}, \gamma^2) \mathbb{P}(X_{i,age} = x_{i,age} | \mu, \sigma^2) \\
&= \prod_{i=1}^N \text{Normal}(Y_i | \beta_0 + \beta_{age}\mu + \sum \beta_j X_{i,j} + \varepsilon_i, \gamma^2) \text{Normal}(X_{i,age} = x_{i,age} | \mu, \sigma^2).
\end{aligned}$$

In Example 2, we discussed how to derive the likelihood of a linear regression model with multiple independent variables because income is likely a factor of several variables. But many times it is very difficult to account for all the possible features in the true data generating process, and our model will therefore have some omitted variables. In other words, there exist quantities that we did not include in our model that may explain variation in our dependent variable. When we assume the omitted quantities are constants and we include a term in our regression for them, we refer to the model as a fixed effects linear regression model. And when we assume the omitted quantities are variable, we refer to the model as a random effects linear regression model. Let us consider an example of a random effects model.

Definition 0.2: Fixed Effects Linear Regression

Let Y_i be our dependent variable for observation i with m independent variables $X_{i,1}, \dots, X_{i,m}$. Our model is known as a **fixed effects linear regression** if it has the following form:

$$\begin{aligned}
& \text{for } i = 1, \dots, n \\
& \quad \varepsilon_i \sim \text{Normal}(\mu, \sigma^2) \\
& \quad \text{for } j = 1, \dots, m \\
& \quad X_{i,j} \sim \mathbb{P}(X_{i,j}|\theta) \text{ (if } X_{i,j} \text{ is variable)} \\
& Y_i | \beta_0, \beta_1, \dots, \beta_j, \theta, X_{i,1}, \dots, X_{i,m} = \beta_0 + \sum_{j=1}^m \beta_j X_{i,j} + \varepsilon_i + \alpha_i.
\end{aligned}$$

Here, α_i represents all the constant variation in Y_i uncaptured by the traditional linear regression model for individual i .

Definition 0.3: Random Effects Linear Regression

Let Y_i be our dependent variable for observation i with m independent variables $X_{i,1}, \dots, X_{i,m}$. Our model is known as a **random effects linear regression** if it has the following form:

$$\begin{aligned}
& \text{for } i = 1, \dots, n \\
& \quad \varepsilon_i \sim \text{Normal}(\mu, \sigma^2) \\
& \quad \text{for } j = 1, \dots, m \\
& \quad X_{i,j} \sim \mathbb{P}(X_{i,j}|\theta) \text{ (if } X_{i,j} \text{ is variable)} \\
& \quad \alpha_i \sim \mathbb{P}(\alpha) \\
& Y_i | \beta_0, \beta_1, \dots, \beta_j, \theta, X_{i,1}, \dots, X_{i,m} = \beta_0 + \sum_{j=1}^m \beta_j X_{i,j} + \varepsilon_i + \alpha_i.
\end{aligned}$$

Here, α_i represents all the variation in Y_i uncaptured by the traditional linear regression model for individual i .

Example 3. Suppose we administer a survey to 100 people asking about their incomes and ages, denoted as Y_i and X_i respectively for individual i . Suppose respondents 1 to 97 answered all questions, persons 98 and 99 only reported age, and person 100 reported only age but not income. In other words, our data set would read as follows:

We also assume that the true data generating process is

$$\begin{aligned}
& \text{for } i = 1, \dots, 100 \\
& \quad \varepsilon_i \sim \text{Normal}(0, \sigma^2) \\
& \quad X_i \sim \text{Normal}(\mu, \gamma^2) \\
& Y_i | \varepsilon_i, X_i = \beta_0 + \beta_1 X_i + \varepsilon_i.
\end{aligned}$$

Because X_i and Y_i come from probability distributions, X_i and Y_i are random variables. So, X_1, \dots, X_{100} and Y_1, \dots, Y_{100} are random variables. However, because X_{98}, X_{99}, Y_{100} are missing from the dataset, these

Table 2: Missing Survey Data Statement

Obs#	Y	X
1	y_1	x_1
2	y_2	x_2
\vdots	\vdots	\vdots
97	y_{97}	x_{97}
98	y_{98}	?
99	y_{99}	?
100	?	x_{100}

are latent variables. Because we have these latent variables that explain variation in our model, this is akin to the random effects linear regression we just introduced.

So, our 2x2 table would read as follows:

	Observed	Unobserved
Variable	$X_1, \dots, X_{97}, X_{100}, Y_1, \dots, Y_{99}$	X_{98}, X_{99}, Y_{100}
Constant		$\beta_0, \beta_1, \sigma^2, \mu, \gamma^2$

As usual, from the 2x2 table, we want to derive the proper likelihood, which is the probability of our observed variables given our constant terms:

$$\text{Likelihood} = \mathbb{P}(X_1 = x_1, \dots, X_{97} = x_{97}, X_{100} = x_{100}, Y_1 = y_1, \dots, Y_{99} = y_{99} | \beta_0, \beta_1, \sigma^2, \mu, \gamma^2).$$

Notice, in the proper likelihood, the latent variables do not exist. But we need the information from those quantities to capture the entire model's process. We can do this by taking advantage of marginalizing distributions (i.e., $\mathbb{P}(A) = \int_B \mathbb{P}(A, B) dB$ for some random variables A and B). So,

$$\begin{aligned} \text{Likelihood} &= \mathbb{P}(X_1 = x_1, \dots, X_{97} = x_{97}, Y_1 = y_1, \dots, Y_{99} = y_{99} | \beta_0, \beta_1, \sigma^2, \mu, \gamma^2) \\ &= \int_{X_{98}} \int_{X_{99}} \int_{Y_{100}} \mathbb{P}(X_1 = x_1, \dots, X_{100} = x_{100}, Y_1 = y_1, \dots, Y_{100} = y_{100} | \beta_0, \beta_1, \sigma^2, \mu, \gamma^2) dX_{98} dX_{99} dY_{100}. \end{aligned}$$

Now, we can take advantage of the independence of these values and write out the joint likelihood as the product of individual likelihoods:

$$\begin{aligned} \text{Likelihood} &= \mathbb{P}(X_1 = x_1, \dots, X_{97} = x_{97}, Y_1 = y_1, \dots, Y_{99} = y_{99} | \beta_0, \beta_1, \sigma^2, \mu, \gamma^2) \\ &= \int_{X_{98}} \int_{X_{99}} \int_{Y_{100}} \mathbb{P}(X_1 = x_1, \dots, X_{100} = x_{100}, Y_1 = y_1, \dots, Y_{100} = y_{100} | \beta_0, \beta_1, \sigma^2, \mu, \gamma^2) dX_{98} dX_{99} dY_{100} \\ &= \int_{X_{98}} \int_{X_{99}} \int_{Y_{100}} \prod_{i=1}^{100} \mathbb{P}(X_i = x_i, Y_i = y_i | \beta_0, \beta_1, \sigma^2, \mu, \gamma^2) dX_{98} dX_{99} dY_{100}. \end{aligned}$$

And now, by Bayes' Rule we can rewrite the joint probability of X_i and Y_i as

$$\begin{aligned}
\text{Likelihood} &= \mathbb{P}(X_1 = x_1, \dots, X_{97} = x_{97}, Y_1 = y_1, \dots, Y_{99} = y_{99} | \beta_0, \beta_1, \sigma^2, \mu, \gamma^2) \\
&= \int_{X_{98}} \int_{X_{99}} \int_{Y_{100}} \mathbb{P}(X_1 = x_1, \dots, X_{100} = x_{100}, Y_1 = y_1, \dots, Y_{100} = y_{100} | \beta_0, \beta_1, \sigma^2, \mu, \gamma^2) dX_{98} dX_{99} dY_{100} \\
&= \int_{X_{98}} \int_{X_{99}} \int_{Y_{100}} \prod_{i=1}^{100} \mathbb{P}(X_i = x_i, Y_i = y_i | \beta_0, \beta_1, \sigma^2, \mu, \gamma^2) dX_{98} dX_{99} dY_{100} \\
&= \int_{X_{98}} \int_{X_{99}} \int_{Y_{100}} \prod_{i=1}^{100} \mathbb{P}(Y_i = y_i | X_i = x_i, \beta_0, \beta_1, \sigma^2) \mathbb{P}(X_i = x_i | \mu, \gamma^2) dX_{98} dX_{99} dY_{100}.
\end{aligned}$$

Since Y_i is a linear combination of two normally distributed random variables (ε_i and X_i), $Y_i \sim \text{Normal}(\beta_0 + \beta_1 \mu, \sigma^2)$. So, we can rewrite the likelihood as

$$\begin{aligned}
\text{Likelihood} &= \mathbb{P}(X_1 = x_1, \dots, X_{97} = x_{97}, Y_1 = y_1, \dots, Y_{99} = y_{99} | \beta_0, \beta_1, \sigma^2, \mu, \gamma^2) \\
&= \int_{X_{98}} \int_{X_{99}} \int_{Y_{100}} \mathbb{P}(X_1 = x_1, \dots, X_{100} = x_{100}, Y_1 = y_1, \dots, Y_{100} = y_{100} | \beta_0, \beta_1, \sigma^2, \mu, \gamma^2) dX_{98} dX_{99} dY_{100} \\
&= \int_{X_{98}} \int_{X_{99}} \int_{Y_{100}} \prod_{i=1}^{100} \mathbb{P}(X_i = x_i, Y_i = y_i | \beta_0, \beta_1, \sigma^2, \mu, \gamma^2) dX_{98} dX_{99} dY_{100} \\
&= \int_{X_{98}} \int_{X_{99}} \int_{Y_{100}} \prod_{i=1}^{100} \mathbb{P}(Y_i = y_i | X_i = x_i, \beta_0, \beta_1, \sigma^2) \mathbb{P}(X_i = x_i | \mu, \gamma^2) dX_{98} dX_{99} dY_{100} \\
&= \int_{X_{98}} \int_{X_{99}} \int_{Y_{100}} \prod_{i=1}^{100} \text{Normal}(\beta_0 + \beta_1 x_i, \sigma^2 | X_i = x_i) \text{Normal}(X_i = x_i | \mu, \gamma^2) dX_{98} dX_{99} dY_{100}.
\end{aligned}$$

Unsupervised Learning

In supervised learning, we observe the outcome that we are aiming to predict. However, there are many use cases in which the outcome of interest is in fact a latent variable. For instance, suppose we run a business and want to find customers who are similar to each other to recommend content; customer similarity is unobserved but random, so the quantity we want to predict is latent. Or, if we are analyzing tweets in specific communities, we may want to know what topics people are interested in and how that changes over time; because the topics are unobserved, they are latent variables. We will consider two unsupervised learning models – one for understanding the interconnectedness of people and another for discovering topics in text data.

Latent Space Models with Network Data

Latent space models are a class of models that allow us to identify the position of different units in an unobserved space by taking advantage of network data. Because the goal of the model is to estimate the latent positions of our units, which are unobserved quantities, these are all unsupervised learning models. The basic intuition relies on the idea that the closer two units are in the latent space, the more likely they are to be connected in a network. By understanding the latent space of the units in our population, we can then identify similar customers in a business for product recommendation, or we can identify which communities in

a network are related, or understand the connectivity patterns in brain networks.

The model assumes that we have N individuals and an adjacency matrix Y such that the entry in row i and column j , $Y_{i,j}$ is 1 if individuals i and j are connected and 0 otherwise. Additionally, for each pair of units i and j , we have observed characteristics about their shared relationship $X_{i,j}^1, X_{i,j}^2, \dots, X_{i,j}^m$ and their unobserved latent spaces Z_i and Z_j . Then, the log-odds of individual i and j having an edge between them is given by

$$\eta_{i,j} = \text{logodds}\mathbb{P}(Y_{i,j} = 1 | X_{i,j}^1, \dots, X_{i,j}^m, Z_i, Z_j) = \alpha + \sum_{k=1}^m \beta_k X_{i,j}^k + |Z_i - Z_j|.$$

So, the general data-generating process for this model can be written as follows:

```
# Sample latent space of each unit  $i$  for  $i = 1, \dots, N$  do
|  $Z_i \sim f_Z(Z)$ 
end
for  $i = 1, \dots, N$  do
| for  $j = 1, \dots, N$  do
| | for  $k = 1, \dots, m$  do
| | | # Sample shared covariates if they follow a probability distribution  $X_{i,j}^k \sim f_{X^k}(X^k)$ 
| | | end
| |  $\eta_{i,j} = \text{logodds}\mathbb{P}(Y_{i,j} = 1 | X_{i,j}^1, \dots, X_{i,j}^m, Z_i, Z_j) = \alpha + \sum_{k=1}^m \beta_k X_{i,j}^k + |Z_i - Z_j|$ 
| | end
| end
end
```

Algorithm 1: General Latent Space Modeling DGP

The reason we use log-odds rather than the raw probability is that the log-odds have a co-domain between $-\infty$ and ∞ while the raw probability has a codomain between 0 and 1. Notice, the right hand side is very similar to the form of a regression: α is akin to the “y-intercept” in the linear regression and β_1, \dots, β_l are the coefficients for each of our covariates. However, α is now the y-intercept for the log-odds and β_1, \dots, β_l are coefficients for our relationship-level covariates. In addition to these differences, we now also include a term for the distance between user i and user j .

In this model, we make no restrictions on whether $X_{i,j}^1, \dots, X_{i,j}^l$ are observed; it may even be the case that some features are observed while others are latent. And furthermore, we make no restriction on whether $X_{i,j}^1, \dots, X_{i,j}^l$ are random or constant. As such, the model can be very flexible.

Notice that in the general model, $\eta_{i,j}$ does not have to necessarily be the same as $\eta_{j,i}$; in other words, the chances of individual i having a connection to individual j is not the same as individual j having a connection to individual i . This may be as a result of X^k . Suppose $X_{i,j}^k$ represents the number of messages on a school’s messaging board that individual i messaged individual j ; in this case, $X_{i,j}^k \neq X_{j,i}^k$, so then $\beta_k X_{i,j}^k \neq \beta_k X_{j,i}^k$; thus, $\eta_{i,j} \neq \eta_{j,i}$. If $\eta_{i,j} \neq \eta_{j,i}$, then our resulting network structure would be a directed network, which means that person i could be related to person j but person j does not have to be related to person i . For example, if we consider a friendship network, student i may think that they are friends with student j but student j does not have to reciprocate that relationship.

Additionally, there are many quantities involved and therefore many different possible conceptualizations possible. For example, there are the case when $\mathbf{X}_{i,j}^1, \dots, \mathbf{X}_{i,j}^l$ are constant or when they are variable or there are some pairwise covariates that are random while others are constant. We also have the cases in which

$\mathbf{Z}_{i,1}, \dots, \mathbf{Z}_{i,m}$ are constant or when they are variable or when some individual-level covariates are random while others are constant. We also have the cases in which $\alpha, \beta_1, \dots, \beta_l$ are variable or constant or a mix of the two. The actual 2x2 and likelihood will therefore be very case-dependent. So let us consider a few examples.

Example 4. Suppose we have twenty students in a classroom and want to understand how similar they are to each other to craft our lesson plans; if students are very similar, then we can have a general teaching plan, but if students are very different, then we should have group assignments so that similar students can learn from one another.

We will model this problem as a latent space model and assume the following data generating process:

```

for  $i = 1, \dots, 20$  do
  for  $j = i + 1, \dots, 20$  do
     $\left( \frac{\eta_{i,j}}{1 - \eta_{i,j}} \right) = \alpha + \beta |x_i - x_j|$ 
     $y_{i,j} | \eta_{i,j} = y_{j,i} | \eta_{j,i} = \text{Bernoulli}(\eta_{i,j})$ 
  end
end

```

Algorithm 2: Latent Space Model DGP: Example 4

Recall from the general latent space model, $\eta_{i,j}$ represents the probability of an edge existing between units i and j , and the actual existence of an edge going from i to j is given by $Y_{i,j}$. Similarly, in this example, we have $\eta_{i,j}$ and $Y_{i,j}$ representing the probability and existence of an edge going from i to j respectively; however, we also assume that $Y_{i,j} = Y_{j,i}$, which means that an edge extends from i to j if and only if an edge extends from j to i . In common language, we assume people who sit closer to each other are more likely to be friends with each other and reciprocate that friendship. Also, because $Y_{i,j} = Y_{j,i}$, including both sets of data in our problem is redundant. Instead, we will only look at the $Y_{i,j}$ for which $i < j$. If we represented this data as a matrix, we would only be looking at the upper triangle, as seen in blue in Table 3.

Table 3: Adjacency Matrix: Example 4

	1	2	3	...	19	20
1	0	$y_{1,2}$	$y_{1,3}$...	$y_{1,19}$	$y_{1,20}$
2	$y_{2,1}$	0	$y_{2,3}$...	$y_{2,19}$	$y_{2,20}$
3	$y_{3,1}$	$y_{3,2}$	0	...	$y_{3,19}$	$y_{3,20}$
\vdots	\vdots	\vdots	\vdots	\ddots	\vdots	
19	$y_{19,1}$	$y_{19,2}$	$y_{19,3}$...	0	$y_{19,20}$
20	$y_{20,1}$	$y_{20,2}$	$y_{20,3}$...	$y_{20,19}$	0

Now, we can write the 2x2 table for this model.

Because $\eta_{i,j}$ for all units i and j are not explicitly from a distribution and none of its inputs (i.e., α, x_i, x_j) follow a distribution $\eta_{i,j}$ is a constant. And because $Y_{i,j}$ follows a Bernoulli distribution, $Y_{i,j}$ is a random variable. We are only given $Y_{i,j}$ and no other data. So, our unknowns are $\alpha, \beta, \eta_{1,2}, \eta_{1,3}, \dots, \eta_{1,20}, \eta_{2,3}, \eta_{2,4}, \dots, \eta_{2,20}, \dots, \eta_{19,20}$ – all of which are constants. So, our 2x2 table would look as follows:

	Observed	Unobserved
Variable	$Y_{1,2}, Y_{1,3}, \dots, Y_{1,20}, Y_{2,3}, Y_{2,4}, \dots, Y_{2,20}, \dots, Y_{19,20}$	
Constant		$\alpha, \eta_{1,2}, \eta_{1,3}, \dots, \eta_{1,20}, \eta_{2,3}, \eta_{2,4}, \dots, \eta_{2,20}, \dots, \eta_{19,20}, \beta$

Recognize that we do not include x_i, x_j in our 2x2 table: this is because $\eta_{i,j}$ and x_i, x_j are deterministically related so if we know $\eta_{i,j}$, we can reconstruct x_i, x_j . And similarly, if we know x_i, x_j , we can reconstruct $\eta_{i,j}$.

Then, our likelihood would be

$$\text{Likelihood} = \mathbb{P}(Y_{1,2} = y_{1,2}, \dots, Y_{1,20} = y_{1,20}, Y_{2,3} = y_{2,3}, \dots, Y_{2,20} = y_{2,20}, \dots, Y_{19,20} = y_{19,20} | \alpha, \beta, \eta_{1,2}, \dots, \eta_{19,20}).$$

Recall that one of the core assumptions of the latent space models is that the existence of an edge between two individuals is independent from all other edges conditioned on the latent space and other constants. So, we can then rewrite the joint likelihood of edges existing as the product of individual likelihoods of edges existing:

$$\begin{aligned} \text{Likelihood} &= \mathbb{P}(Y_{1,2} = y_{1,2}, \dots, Y_{1,20} = y_{1,20}, Y_{2,3} = y_{2,3}, \dots, Y_{2,20} = y_{2,20}, \dots, Y_{19,20} = y_{19,20} | \alpha, \beta, \eta_{1,2}, \dots, \eta_{19,20}) \\ &= \prod_{1 \leq i < j \leq 20} \mathbb{P}(Y_{i,j} = y_{i,j} | \alpha, \beta, \eta_{i,j}). \end{aligned}$$

Because $Y_{i,j}$ follows a Bernoulli distribution with success probability $\eta_{i,j}$,

$$\mathbb{P}(Y_{i,j} = y_{i,j} | \alpha, \beta, \eta_{i,j}) = [\mathbb{P}(Y_{i,j} = 1 | \alpha, \beta, \eta_{i,j})]^{y_{i,j}} [\mathbb{P}(Y_{i,j} = 0 | \alpha, \beta, \eta_{i,j})]^{1-y_{i,j}}.$$

Additionally, because $\mathbb{P}(Y_{i,j} = 1) = \eta_{i,j}$, we can further simplify $\mathbb{P}(Y_{i,j})$ to

$$\mathbb{P}(Y_{i,j} = y_{i,j} | \alpha, \beta, \eta_{i,j}) = [\eta_{i,j}]^{y_{i,j}} [1 - \eta_{i,j}]^{1-y_{i,j}}.$$

Example 5. Now, let us consider an example in which $\eta_{i,j}$ actually follows a probability distribution. Suppose we have the same DGP as before, except $x_i, \overset{\text{iid}}{\sim} \text{Normal}(0, \sigma^2)$:

```

for  $i = 1, \dots, 20$  do
  |  $X_i \sim \text{Normal}(\mu, \sigma^2)$ 
end
for  $i = 1, \dots, 20$  do
  | for  $j = i + 1, \dots, 20$  do
  |   |  $\eta_{i,j} = \log \frac{\mathbb{P}(Y_{i,j}=1 | \alpha, \beta, X_i, X_j)}{\mathbb{P}(Y_{i,j}=0 | \alpha, \beta, X_i, X_j)} = \alpha + \beta |X_i - X_j|$ 
  |   end
  | end
end

```

Algorithm 3: Latent Space Model DGP: Example 5

Notice, we rewrote the same likelihood but simply with a different form. Here, we explicitly write out what $\eta_{i,j}$ represents: the log-odds of $Y_{i,j}$ being connected. That means that the success probability of $Y_{i,j}$ being 1 would be given by $\frac{1}{1+e^{\alpha+\beta|X_i-X_j|}}$. Let us find the likelihood of this new problem.

Similar to Example 4, we assume the only data we are given is the adjacency matrix; and because we assume $Y_{i,j} = Y_{j,i}$, we only need data on $Y_{i,j}$ when $1 \leq i < j \leq 20$. So, our 2x2 table would be the following:

$$\begin{array}{c} \text{var_table} \\ \{Y_{1,2}, Y_{1,3}, \dots, Y_{1,20}, Y_{2,3}, Y_{2,4}, \dots, Y_{2,20}, \dots, Y_{19,20}\} \end{array} \begin{array}{c} \{X_1, \dots, X_{20}\} \\ \{\alpha, \beta, \mu, \sigma^2\} \end{array}$$

Again, because $\eta_{i,j}$ and X_i, X_j are deterministically related, we only need to use one or the other in our 2x2 table. In this example, we will now use X_1, \dots, X_{20} instead for our calculations. Then, the proper likelihood is given by:

$$\text{Likelihood} = \mathbb{P}(Y_{1,2} = y_{1,2}, \dots, Y_{19,20} = y_{19,20} | \alpha, \beta, \mu, \sigma^2).$$

However, because we now have latent variables, we need to consider X_1, \dots, X_{20} as well: to do this, we must recognize the proper likelihood is the marginal of the complete likelihood. So, we can integrate out X_1, \dots, X_{20} from the joint complete likelihood and obtain the proper likelihood:

$$\text{Likelihood} = \int_{X_1} \dots \int_{X_{20}} \mathbb{P}(Y_{1,2} = y_{1,2}, \dots, Y_{19,20} = y_{19,20}, X_1 = x_1, \dots, X_{20} = x_{20} | \alpha, \beta, \mu, \sigma^2) dx_1 \dots dx_{20}.$$

Now, we can take advantage of Bayes' Rule and rewrite the likelihood as

$$\text{Likelihood} = \int_{X_1} \dots \int_{X_{20}} \mathbb{P}(Y_{1,2} = y_{1,2}, \dots, Y_{19,20} = y_{19,20} | X_1 = x_1, \dots, X_{20} = x_{20}, \alpha, \beta) \mathbb{P}(X_1 = x_1, \dots, X_{20} = x_{20} | \mu, \sigma^2) dx_1 \dots dx_{20}.$$

Recall the conditional independence assumption of the latent space model. So, that means that we can further simplify the joint complete likelihood to be the product of individual complete likelihoods:

$$\text{Likelihood} = \int_{X_1} \dots \int_{X_{20}} \prod_{1 \leq i < j \leq 20} \mathbb{P}(Y_{i,j} = y_{i,j} | X_i = x_i, X_j = x_j, \alpha, \beta) \mathbb{P}(X_i = x_i, X_j = x_j | \mu, \sigma^2) dx_1 \dots dx_{20}.$$

But since $X_i \stackrel{\text{iid}}{\sim} \text{Normal}(\mu, \sigma^2)$, $\mathbb{P}(X_i = x_i, X_j = x_j | \mu, \sigma^2) = \mathbb{P}(X_i = x_i | \mu, \sigma^2) \mathbb{P}(X_j = x_j | \mu, \sigma^2)$. So,

$$\begin{aligned} \text{Likelihood} &= \int_{X_1} \dots \int_{X_{20}} \prod_{1 \leq i < j \leq 20} \mathbb{P}(Y_{i,j} = y_{i,j} | X_i = x_i, X_j = x_j, \alpha, \beta) \mathbb{P}(X_i = x_i | \mu, \sigma^2) \mathbb{P}(X_j = x_j | \mu, \sigma^2) dx_1 \dots dx_{20} \\ &= \int_{X_1} \dots \int_{X_{20}} \prod_{1 \leq i < j \leq 20} \text{Bernoulli}(Y_{i,j} = y_{i,j} | X_i = x_i, X_j = x_j, \alpha, \beta) \text{Normal}(X_i = x_i | \mu, \sigma^2) \text{Normal}(X_j = x_j | \mu, \sigma^2) dx_1 \dots dx_{20} \end{aligned}$$

Let us break this into more tractable pieces. Since $Y_{i,j}$ follows a Bernoulli distribution with success

$$\frac{1}{1 + e^{\alpha + \beta |X_i - X_j|}},$$

$$\mathbb{P}(Y_{i,j} = y_{i,j} | \alpha, \beta, X_i = x_i, X_j = x_j) = \left[\frac{1}{1 + e^{\alpha + \beta |X_i - X_j|}} \right]^{y_{i,j}} \left[\frac{e^{\alpha + \beta |X_i - X_j|}}{1 + e^{\alpha + \beta |X_i - X_j|}} \right]^{1 - y_{i,j}}.$$

Now, since $X_i, X_j \sim \text{Normal}(\mu, \sigma^2)$,

$$f_X(x_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}.$$

So,

$$f_X(x_i|\mu, \sigma^2)f_X(x_j|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x_j-\mu}{\sigma}\right)^2},$$

which we can then simplify to

$$\frac{1}{2\pi\sigma^2} e^{-\frac{1}{2\sigma^2}((x_i-\mu)^2+(x_j-\mu)^2)}.$$

So, putting all the pieces together,

$$\begin{aligned} \text{Likelihood} &= \mathbb{P}(Y_{1,2} = y_{1,2}, \dots, Y_{19,20} = y_{19,20} | \alpha, \beta, \mu, \sigma^2) \\ &= \int_{X_1} \dots \int_{X_{20}} \prod_{1 \leq i < j \leq 20} \left[\frac{1}{1 + e^{\alpha + \beta|X_i - X_j|}} \right]^{y_{i,j}} \left[\frac{e^{\alpha + \beta|X_i - X_j|}}{1 + e^{\alpha + \beta|X_i - X_j|}} \right]^{1 - y_{i,j}} \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2\sigma^2}((x_i-\mu)^2+(x_j-\mu)^2)} dx_1 \dots dx_{20}. \end{aligned}$$