Chapter 2

Data

In general, we might think of "data" as measurements of some physical process at discrete points in time, or at discrete points in space, for example. The data can therefore depend on some temporal process (a *time series*), like the stock values of a certain company during the day, the oil pressure in a running engine, the atmospheric temperature at a given spot on the surface of the earth, the population of rabbits on an island, etc. The data may not depend on time; for example, the pixel values on a photograph, the analysis of a document (perhaps by using a word count), the answers to a questionnaire, the results of a blood test, etc.

In looking at a data set for the first time, you should be asking yourself: What knowledge am I trying to extract from this data? It is in this sense that data carries *information*, and it is our task to figure out what that information is[1]. Our task is made harder by the fact that, in general, the data will generally be represented in a form that hides the information (if not, our task is greatly simplified!). That is, given set of p data points in $\mathbb{R}^n$,

$$\left\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}, \ldots, \mathbf{x}^{p}\right\}$$

Suppose we have some invertible, continuous function F. We will say that the new data set,

$$\mathbf{y}^{(k)} = F(\mathbf{x}^{(k)})$$

is similar to the data set $\mathbf{x}$ in that the information carried by the data will not change (Given the data set in $\mathbf{y}$, we could re-construct the data set in $\mathbf{x}$). Commonly, the function F will be a change of basis.

2.1 Representations of Data, Part I

In linear algebra, we learned techniques that allowed us to translate vectors from one basis to another. For example, suppose we have a set of data, $\left\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}\right\}$,

[1] You may think we're heading into an area known as *information theory*, but we'll only be looking at it superficially.

where $\mathbf{x}^{(i)} \in \mathbb{R}^{10}$. The standard way of representing this data is to use the standard basis, $\mathbf{e}_1, \ldots, \mathbf{e}_{10}$:

$$\mathbf{x}^{(i)} = \sum_{k=1}^{10} x_k^i \mathbf{e}_k$$

where x_k^i is the k^{th} component of vector i. But this seems like overkill- do you really need 10 basis vectors to describe three data points? Here's a better representation:

$$\begin{aligned}
\mathbf{x}^{(1)} &= 1 \cdot \mathbf{x}^{(1)} + 0 \cdot \mathbf{x}^{(2)} + 0 \cdot \mathbf{x}^{(3)} \\
\mathbf{x}^{(2)} &= 0 \cdot \mathbf{x}^{(1)} + 1 \cdot \mathbf{x}^{(2)} + 0 \cdot \mathbf{x}^{(3)} \\
\mathbf{x}^{(3)} &= 0 \cdot \mathbf{x}^{(1)} + 0 \cdot \mathbf{x}^{(2)} + 1 \cdot \mathbf{x}^{(3)}
\end{aligned}$$

In this setting, $\mathbf{x}^{(1)}$ is represented as $(1,0,0)$, $\mathbf{x}^{(2)}$ is $(0,1,0)$, and $\mathbf{x}^{(3)}$ is $(0,0,1)$. We are **reducing the dimension** of the set from $\mathbb{R}^{10}$ to $\mathbb{R}^3$. If we were to further assume that $\mathbf{x}^3$ is actually $2\mathbf{x}^{(1)} - 3\mathbf{x}^{(2)}$, then the dimension can be reduced even further:

$$\mathbf{x}^{(1)} \to (1,0), \quad \mathbf{x}^{(2)} \to (0,1), \quad \mathbf{x}^{(3)} \to (2,-3)$$

A key idea here is that a data set spans itself.

In general, if we have a basis, $\mathcal{B} = \{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ for some $k-$dimensional vector space, then any point in that space can be written as:

$$\mathbf{x} = \sum_{j=1}^{k} \alpha_i \mathbf{v}_i \doteq B[x]_b$$

where B is the matrix whose columns are the basis vectors, $[\mathbf{v}_1 \ , \ldots, \mathbf{v}_k]$ and $[x]_b$ is the coordinate vector of $\mathbf{x}$ with respect to that basis. If we have two sets of bases, $\mathcal{B}, \mathcal{C}$, then we can define the change of coordinates via the relations:

$$\mathbf{x} = B[x]_b \; \mathbf{x} = C[x]_c \Rightarrow B[x]_b = C[x]_c$$

so that we can determine either $[x]_b, [x]_c$ given the other set of coordinates.

So, there are an infinite number of choices for the basis. The question of which to choose will be left to consider later on.

A related notion is that a matrix of numbers has a dual nature in that if we are just presented with an $m \times n$ matrix, we do not know if it represents m data points in $\mathbb{R}^n$ or n data points in $\mathbb{R}^m$. We will consider this duality when we introduce the Singular Value Decomposition (SVD).

2.2 Representations of Data, Part II

For data that is a time series (that is, the data is given to us as a string of real numbers), it is a common practice to *embed* the time series to $\mathbb{R}^k$ in the following way:

Given $x_1, x_2, x_3, x_4, \ldots$, a sequence of numbers, the following represents the data embedded in $\mathbb{R}^2$ via lag 2:

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \begin{bmatrix} x_2 \\ x_3 \end{bmatrix}, \begin{bmatrix} x_3 \\ x_4 \end{bmatrix}, \ldots$$

or we could embed the sequence in $\mathbb{R}^3$ via lag 3:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \begin{bmatrix} x_2 \\ x_3 \\ x_4 \end{bmatrix}, \begin{bmatrix} x_3 \\ x_4 \\ x_5 \end{bmatrix}, \ldots$$

and so on. We'll discuss this technique in depth later on.

Why do such a thing? To plot a sequence of real numbers, we normally plot the index i along the horizontal axis, with the value of x_i along the vertical. For example, the data may be timed samples from a sine function,

$$x_t = \sin(t/3)$$

If we plot the ordered pairs (t, x_t), the graph of the function leaves our plot since time is unbounded.

On the other hand, if x_t is bounded for all time (in our example, $|x_t| < 1$), then plotting the embedded pairs (x_i, x_{i+1}) gives us the full graph of the function (a circle or an ellipse for our example).

2.3 Representing Functions instead of Data

As we have seen, a given set of data can be viewed in terms of some discrete sampling of points,

$$f(t_0), f(t_1), f(t_2), \ldots, f(t_N)$$

but it is interesting to consider other types of representations.

Consider the following example from linear algebra: The functions $v_1(x) = 1$, $v_2(x) = x$, $v_3(x) = x^2$ are linearly independent (you might review what that means in terms of functions), so they form a basis for their spanning set. (a) Write $f(x) = 3 - x^2$ in terms of the basis, and (b) Write the function corresponding to $3v_1 - 2v_2 + v_3$.

SOLUTION: In the first case, the function can be expressed as: $3 - x^2 = 3 \cdot 1 - 1 \cdot x^2 = 3v_1 - v_3$. As a set of coordinates, $f = (3, 0, -1)$. In the second example, substitute the functions in to get $3 - 2x + x^2$.

Notice that in this case, the entire **function** is represented as a set of coordinates.

As another example, the underlying function might be represented by a sum of sines and cosines:

$$f(t) = a_0 + \sum_{k=1}^{m} a_k \sin(kt) + b_k \cos(kt)$$

in which case, f could be represented by its *frequency* content (versus its *temporal* content). That is, I could define f by telling you what frequencies f consists of (that is, I could give you numbers for a_i, b_i) , rather than telling you what f is at various times. We will consider this more in depth in the section on Fourier analysis.

In fact, this is a special case of a more abstract setting. Given that f is some linear combination of functions $\phi_k(t)$,

$$f(t) = \sum_{k=1}^{m} b_k \phi_k(t)$$

f could be defined by stating the weights of the linear combination, b_k. An example we're familiar with is the Taylor series for f based at $t = a$. In this case, $\phi_k(t) = (t-a)^k$, and:

$$f(t) = \sum_{k=0}^{\infty} b_k (t-a)^k$$

where we know that $b_k = \frac{1}{k!} f^{(k)}(a)$, where $f^{(k)}$ is the k^{th} derivative of f, and $f^{(0)} = f$.

From linear algebra, we know that $C[a,b]$ is an infinite dimensional vector space, and so requires an infinite number of basis vectors. In our previous examples, the functions ϕ_k are simply the basis vectors, and the b_k are the weights in that basis. Once we discretize a function to only n data points, the set of functions over those points becomes an $n-$ dimensional space- that is, there are only $n-$ parameters that are required to define any function over those $n-$ points. In fact, this space is isomorphic to $\mathbb{R}^n$ with the standard basis.

The representations given so far are called *Linear Models*. They are called that because the model parameters are simply the weights of a linear combination (but notice that the basis functions themselves are *not* linear!). We will describe some nonlinear models later on.

2.4 Summary

In this chapter, we have discussed the nature of representing data, and the basic types of representation available to us. In ambiguous data, the choice of a basis may carry unintended consequences, as we will discuss in the exercises.

2.5 Exercise Set

1. (Matlab Exercise) Write a Matlab function that will input a time series and output a lagged matrix. The input parameters should include the lag length and a value representing a skip. Some samples of what your code should do follow, where x is a vector of integers from 1 to 10.

- Function call: `A=lagmatrix(x,4,2)`
- Output:

$$A = \begin{bmatrix} 1 & 3 & 5 & 7 \\ 2 & 4 & 6 & 8 \\ 3 & 5 & 7 & 9 \\ 4 & 6 & 8 & 10 \end{bmatrix}$$

- Function call: `A=lagmatrix(x,2,3)`
- Output:

$$A = \begin{bmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \end{bmatrix}$$

2. Suppose you are told that $1, 3, 5, 7, \ldots$ belongs to class A, and $2, 4, 6, 8, \ldots$ belongs to class B. What information is being carried- what is the underlying process that is producing this data? What does 1.5 produce?

3. While replacing a vector by its coordinate is dimensionality reducing, in some problems it is appropriate to look at expanding the dimension to get what is called a sparse representation. That is, the data that we are presented with is the output of some projection, and we wish to find a higher dimensional representation. Some examples:

   - "Shape from Shading" is where we find a three dimensional model of a shape from a two dimensional representation. For example, find the shape of the clown's face from his photo.
   - "The Cocktail Party Problem". It is a well known phenomenon that, if we are in a large room full of people speaking, we can isolate the sound of the person speaking to us. A particular instance of this problem: Given a tape recording of an orchestra, for example, is it possible to isolate each instrument's sound?

   In terms of what we know from linear algebra, why is this such a harder problem that finding a low dimensional representation?

4. Suppose that some geologists have been studying some different types of rocks in a certain area. Each rock has certain characteristics of interest- for example, color, size, weight, and other compositional items. They give you a non-numerical data set of alphabetic characters, where each symbol represents a characteristic. How do you translate those alphabetical characters into numbers that can be analyzed on a computer[2]?

5. What is the smallest number of basis vectors needed to represent the data:

$$\mathbf{x}_1 = \begin{bmatrix} 2 \\ -1 \\ 2 \\ 0 \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} 2 \\ -1 \\ 2 \\ 3 \end{bmatrix}, \mathbf{x}_3 = \begin{bmatrix} -1 \\ \frac{1}{2} \\ -1 \\ 0 \end{bmatrix}, \mathbf{x}_4 = \begin{bmatrix} 6 \\ -3 \\ 6 \\ 9 \end{bmatrix}$$

[2]Caution: Translation into numbers has certain (unintended?) implications- If we just perform a straightforward conversion to the numbers from 1-26, we are bringing along a metric. That is, we are saying that an "A" is closer to "B" than to "Z".

Using this basis, write down the low-dimensional representations for each data point, $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$.

6. This series of exercises is designed to get you to consider what happens when we put a time series into $\mathbb{R}^2$ or $\mathbb{R}^3$.

   Let $y = \sin(x)$, where x is taken as a sampling of the time interval from $-\pi$ to π. Type the following script in Matlab:

```
x=linspace(-pi,pi,3000);  %Get 3000 samples evenly spaced
y=sin(x);
for j=2:20
  plot(y(1:j-1),y(j:end))
  pause
end
```

   (a) Before beginning, show (graphically) that: $-\cos(t) = \sin\left(t + \frac{\pi}{2}\right)$ and that $-\sin(t) = \sin(t + \pi)$.
   (b) Note that the effect of this code is to plot the series y at time t against the series y at time $t + \tau$. Can you predict what will happen? (HINT: Consider what will happen if $\tau = 0$, $\tau = \frac{\pi}{2}$ and $\tau = \pi$ using the previous part).
   (c) Run the code and write up your observations: What happens to the curve as j changes?
   (d) Try changing y to $\cos(x)$. Any changes?
   (e) Try changing x: `x=linspace(0,100*pi,10000)`, and change y to `y=sin(x)+sin((1/sqrt(2))*x)`, and re-run the code. What happens now?

7. The *dimension* of a set varies with the definition used. We explore a couple of these below:

   - We will define the Euclidean Dimension (D_e) of a set S to be the (smallest) number of basis vectors needed to describe the given set. This is a global quantity, and is the definition we used in linear algebra.
   - The Topological Dimension (D_T) of a set S is something that is local in nature- that is, the topological dimension is something that is determined point by point. To understand this definition, recall that an "$\epsilon-$ neighborhood" (in our usual space R^n) about a point x is the set of points y so that $\|x - y\| < \epsilon$.
     - A set has topological dimension 0 if every point has arbitrarily small neighborhoods whose boundaries do not intersect the set.
     - A set S has topological dimension k if each point in S has arbitrarily small neighborhoods whose boundaries meet S in a set of topological dimension $k - 1$, and k is the least nonnegative integer for which this holds.

(a) Suppose S is a circle in the plane. What is D_e, D_T?

(b) Do the same for the sphere ($\|x\| = 1$) in three dimensions.

(c) Suppose S is the letter "X" in the plane. What is D_e, D_T?

(d) Give some examples of sets S where $D_e = D_T$.

(e) Consider the set S in the plane that consists of a filled in square with a line segment attached to one side. Discuss any problems you might have with computing D_T. In determining D_T, be sure to read its definition carefully.

Side Remark: The topological dimension is always an integer (by the way we've defined it). There is also such a thing as a fractal dimension- In this case, the dimension of a set can be a fraction (for example, a dimension of $3/2$ is "more than a line, but does not fill a plane".). This is discussed in a class on fractal geometry.