Lab 1A: Data, Models and Decisions

Juan Elenter, Ignacio Hounie and Alejandro Ribeiro*

January 18, 2024

1 Systems

A system is an entity that processes some inputs to produce some outputs. In engineering practice, inputs and outputs are often quantified and for this reason we denote inputs with a variable x, outputs with a variable y, and the system itself is represented by the function $f(\cdot)$ that codifies the relationship $y = f(x)$. A schema is shown in Figure 1.

This definition is vague on purpose because it is intended to maintain generality. Pretty much any task can be codified as the action of a system. For instance, when we hear a word and understand its meaning we are acting as a system that takes a time varying sequence of air pressures as inputs and produces a categorical representation as an output. When we read a digit, we are a system that takes light patterns as inputs and produces numbers as outputs – a number being the common property of finite sets that can be related with a bijection. When we watch and rate a movie, we are a system that takes movies as inputs and produces ratings as outputs. Driving a car around a circuit requires a driver that takes as input the desired trajectory and produces a sequence of acceleration inputs that results in the car following the circuit. The driver is a system.

Just as important, this definition of a system is vague because we can keep it vague and yet make it useful. Despite their significant differences all of the systems that we describe above can be represented by the schema

*In alphabetical order.

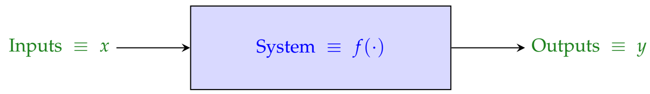

Figure 1. Systems. A system $f(\cdot)$ is an entity that processes inputs x to produce outputs $y = f(x)$. This definition is vague so as to yield a wide range of specific instances. However general, it abstracts properties that make it useful.

in Figure 1. They all process some input – audio, light, movies, or target trajectories – to produce some output – concepts, numbers, ratings, or acceleration sequences. They are, in fact, four systems that we will study in this course.

2 Artificial Intelligence

A first definition of an artificial intelligence (AI) is that of a system that mimics the input-output relationship of a natural system; see Figure 2. When the natural system is presented with the input x it responds by producing the output y. When the AI is presented with the same input x it responds by producing the output $\hat{y}$. If the AI is a good AI, the outputs it produces in response to a given input are similar to the outputs produced by the natural system.

Having an AI is useful because it can be used in lieu of the natural system. For instance, suppose that the input x represents an image of a digit. The natural system is a standard human that looks at this image and reads the number y that it represents. If the AI can successfully mimic a human by spitting numbers $\hat{y}$ that are equal to the numbers recognized by a human reader we can use it in lieu of humans to recognize digits. This is good because the AI frees humans from the drudgery of reading digits. AI's have been in use since the 1990's to recognize digits in checks.

We must point out that the schema in Figure 2 is a flawed definition of AI. A 100 gram tungsten ball dropped from 1 meter mimics quite well any other tungsten ball dropped from 1 meter. Some would say that the definition is less flawed if we add the restriction that the emulated natural system is intelligent but this begs the question of what it means for a natural system to be intelligent. Besides, an artificial intelligence can

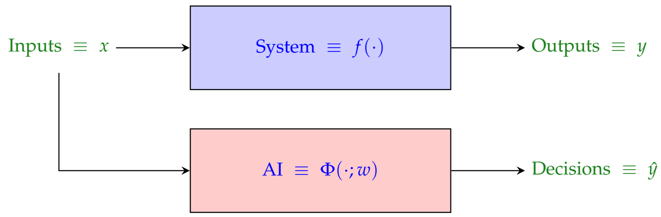

Figure 2. Artificial Intelligence. An artificial intelligence (AI) is a system that mimics the input-output relationship of a natural system. This is a flawed definitions which is nonetheless a good operational definition of the practice of AI.

still be useful if the system that it imitates is not intelligent.

Flawed or not, the most important fact is that Figure 2 is a good operational definition which captures well the current practice of AI.

2.1 Data, Models, and Decisions

To design an AI system the first step is to acquire *data*. This is typically in the form of a set of N input-output pairs (x_i, y_i). We call this collection of examples the training set.

The next step is the selection of a *model*. This is usually in the form of a postulated relationship between inputs and outputs. This is written in Figure 2 as the function $\Phi(\cdot, w)$. The choice of model follows from our knowledge and understanding of the system. For example, convolutional neural networks (CNNs) have invariance and stability properties that make them adequate to process times series and images – as we will see in Labs 2 and 4. The choice of model also follows from accumulated empirical evidence of which models are known to work for specific kinds of systems.

As per Figure 2, when the AI is presented with the input x it produces the output $\hat{y} = \Phi(x, w)$. This output is the AI's estimate or prediction of the actual output y that the system produces when presented with input x. We also say that $\hat{y}$ is the AI's *decision*.

2.2 Machine Learning

In the AI's decisions $\hat{y} = \Phi(x, w)$ the variable w is a parameter that has to be chosen. To choose this parameter we introduce a metric $\ell(y, \hat{y})$ to compare predicted outputs $\hat{y}$ with actual outputs y. We then search for the parameter w that minimizes this loss over the given set of input output pairs,

$$w^* = \underset{w}{\operatorname{argmin}} \sum_{i=1}^{N} \ell\Big(y_i, \Phi(x_i, w) \Big). \tag{1}$$

We call this formulation a supervised machine learning problem (ML). The process of finding the parameter w^* that minimizes the loss averaged over the available data is called training.

We call (1) a supervised learning problem because the AI is given examples of inputs x_i and their corresponding outputs y_i. Alternatively, we may be given example inputs and a cost function $c(\cdot)$ that assesses the merit of the AI decision $\Phi(x_i, w)$. In this case we formulate the unsupervised ML problem,

$$w^* = \underset{w}{\operatorname{argmin}} \sum_{i=1}^{N} c\Big(\Phi(x_i, w) \Big). \tag{2}$$

We must point out that (1) and (2) are controversial definitions of learning. There is nothing in the specifications to represent understanding, although some people argue that sufficiently complex imitation *is* understanding. As we did with the definition of AI, we remain agnostic to this discussion. Equations (1) and (2) are good operational definitions of learning which capture well the current practice of ML.

3 Admissions at the University of Pennsylvania

Let us pretend that we are tasked with designing a system to make admission decisions at the University of Pennsylvania (Penn). In order to design this system we need to acquire data, choose a model, and train it to make admission decisions.

ID	HS GPA	SAT	Gender	Penn GPA
A41675	3.93	1,540	F	3.67
CE58D3	3.79	1,540	M	3.53
C38C00	3.93	1,560	F	3.80
CFA232	3.90	1,570	M	3.53
B57BF6	3.95	1,500	F	3.67
CA694D	3.76	1,520	M	3.40

Figure 3. Grade point average (GPA) and Scholastic Assessment Test (SAT) data samples.

3.1 Data

Figure 3 shows data that we have available to make admission decisions. For a collection of *former* students we have access to their high school (HS) grade point average (GPA), their Scholastic Assessment Test (SAT) scores, their gender and their Penn GPA.

The table shows five representative examples, but we have data for a total of 600 students. Figure 4 shows plots in which the horizontal axes are high school GPAs or SAT scores and the vertical axes are Penn GPAs. These plots show that high school GPA is predictive of Penn GPA. Although there is significant variation we can see that higher high school GPA corresponds with higher Penn GPA. This indicates that high school GPAs are useful information for admission decisions as they can predict with some accuracy the GPA that an admitted student may attain at Penn. We can squint and see that the same is more or less true of SAT scores, although the correlation between SAT scores and Penn GPA is weaker.

Task 1 Follow this link to download the GPA and SAT score data. Reproduce the plots in Figure 4. ■

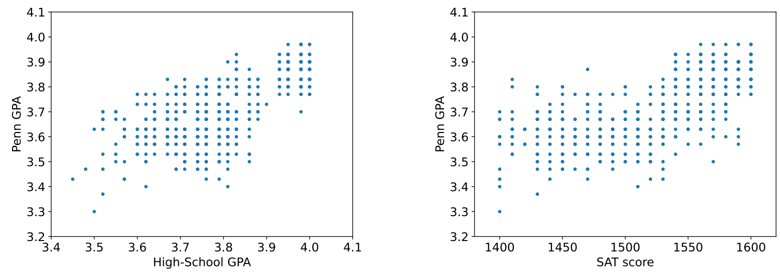

Figure 4. Penn GPA plotted with respect to high school GPA and SAT scores.

3.2 System

To make admission decisions we interpret Penn as the system shown in Figure 5. This system takes as inputs high school GPA, SAT scores and gender information of a student and produces as an output the Penn GPA of the corresponding graduate. If we define the input data as a vector $x = [\text{HS GPA}; \text{SAT}; \text{Gender}]$ and we denote the output as $y = \text{Penn GPA}$ we can represent this system as the function,

$$\text{Penn GPA} \;=\; y \;=\; P(x) \;=\; P \begin{bmatrix} \text{HS GPA} \\ \text{SAT} \\ \text{Gender} \end{bmatrix} . \tag{3}$$

Notice that this is a poor representation of Penn. Incoming students are much more than their gender, High School GPA, and SAT scores. Penn graduates are much more than their Penn GPAs and the institution itself does much more to a high school graduate than transforming their High School GPA and SAT scores into a Penn GPA. This is just one aspect of the whole system on which we are choosing to focus. The distinction between what a system is and what an engineer chooses to say that a system is warrants some discussion that we undertake in Section 4.

3.3 Model and Decisions

To make admission decisions we leverage the system of Section 3.2 and the data of Section 3.1 to design an AI model that *predicts* the Penn GPA of prospective students.

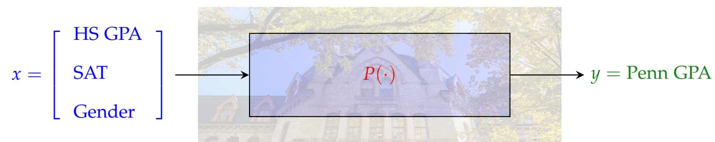

Figure 5. The University of Pennsylvania.

To make matters simpler let us begin by ignoring SAT scores and gender and attempt predictions based on high school GPAs. This means that the system in (4) is replaced by the system

$$\text{Penn GPA} \;=\; y \;=\; P(x) \;=\; P(\,\text{HS GPA}\,). \tag{4}$$

The function $P(x)$ is the true effect of Penn on scholastic accomplishment. This is information that becomes available after the fact. When a student graduates Penn, we have access to their high school GPA x and their Penn GPA y.

Penn GPA predictions are to be made prior the fact. Before a student attends Penn we want to estimate their Penn GPA based on their high school GPA x. We *choose* to postulate a linear relationship and make predictions of the form

$$\hat{y} = \alpha x. \tag{5}$$

In (5), $\hat{y}$ is a *prediction* of the *true* Penn GPA y that will be available after the fact. The coefficient α is to be determined with the goal of making predictions $\hat{y}$ close to to actual Penn GPA y. We can then use Penn GPA predictions to make admission *decisions.*

3.3.1 Least Squares Estimation

To determine a proper value for the coefficient α in (5) we utilize the data we have available on the scholastic performance of past students. Use N to denote the total number of available data points. Introduce a subindex i to differentiate past students so that the pair (x_i, y_i) denotes the high school GPA and Penn GPA of student i. For these students we can make GPA *predictions* $\hat{y}_{=}\alpha x_i$. For a given coefficient α we define the

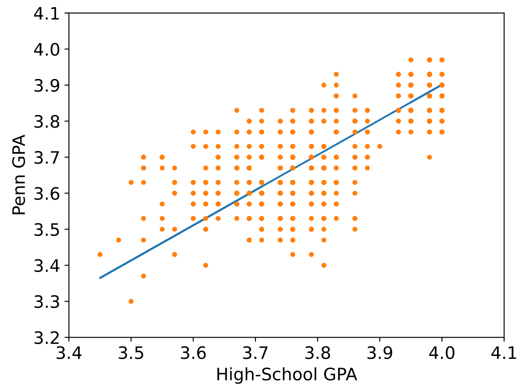

Figure 6. Linear minimum mean squared error (MMSE) prediction of Penn GPA from high school GPA.

mean squared error (MSE),

$$\text{MSE}(\alpha) \;=\; \frac{1}{N}\sum_{i=1}^{N}\Big(y_i - \hat{y}_i\Big)^2 \;=\; \frac{1}{N}\sum_{i=1}^{N}\Big(y_i - \alpha x_i\Big)^2. \tag{6}$$

The mean squared error $\text{MSE}(\alpha)$ measures the predictive power of coefficient α. The quantity $(y_i - \hat{y}_i)^2$ is always nonnegative and indicates how good the predicted GPA $\hat{y}_i$ is to the true GPA y_i. The MSE averages this metric over all students. It follows that a natural choice for α is the value that makes the MSE smallest. We therefore define the optimal coefficient

$$\alpha^* \;=\; \operatorname*{argmin}_{\alpha} \frac{1}{2}\text{MSE}(\alpha) \;=\; \operatorname*{argmin}_{\alpha} \frac{1}{2N}\sum_{i=1}^{N}\Big(y_i - \alpha x_i\Big)^2, \tag{7}$$

and proceed to make Penn GPA predictions as $\hat{y} = \alpha^* x$ [cf. (5)]. This GPA predictor is called the linear minimum mean squared error (MMSE) prediction. This is because the predictor is the linear function that minimizes the MSE.

Task 2 Prove that the MMSE estimator coefficient α^* defined in (7) is given by the expression

$$\alpha^* = \sum_{i=1}^{N} x_i y_i \Bigg/ \sum_{i=1}^{N} x_i^2. \tag{8}$$

Compute α^* for the data loaded in Task 1. Plot the Penn GPA with respect to HS GPA and superimpose the prediction line $\hat{y} = \alpha^* x$. This plot is shown in Figure 6.

In Task 2 we make predictions of the Penn GPA of students that have graduated Penn. Predicting Penn GPAs of past students is unnecessary given that we already know their true GPAs. Our motivation for solving this unnecessary problem is to determine the coefficient α^* that we can use to make predictions $\hat{y} = \alpha^* x$ of students that have not yet attended Penn – for which x is available but y is not. The effectiveness of this prediction depends on the extent to which the past is a good representation of the future.

It is germane to emphasize that in Task 2 we are using something we know – the GPA of former students – to answer a new question – the GPA of a prospective student. However primitive, this is a form of intelligence.

3.3.2 Root Mean Squared Error

We evaluate the merit of α^* with the root mean squared error (RMSE)

$$\text{MSE}(\alpha) \;=\; \sqrt{\text{MSE}(\alpha)} \;=\; \left[\frac{1}{N}\sum_{i=1}^{N}\Big(y_i - \alpha x_i\Big)^2\right]^{1/2}. \tag{9}$$

The reason we use the RMSE to evaluate the merit of α^* instead of the MSE is that the RMSE has the same units. It is easier to interpret than the MSE. Since the difference between the two is just a square root function, the coefficient α^* that minimizes the MSE also minimizes the RMSE.

Task 3 Compute the RMSE of α^* and comment on the quality of the Penn GPA predictions. ■

You should observe that the RMSE is 0.094. We can think of this number as the accuracy of our Penn GPA predictions. This number seems to imply that our Penn GPA predictions are quite accurate because Penn GPAs can range from 0 to 4. However, the actual range of Penn GPAs observed in the dataset is between 3.3 and 4.0. In a variable whose range spans 0.7

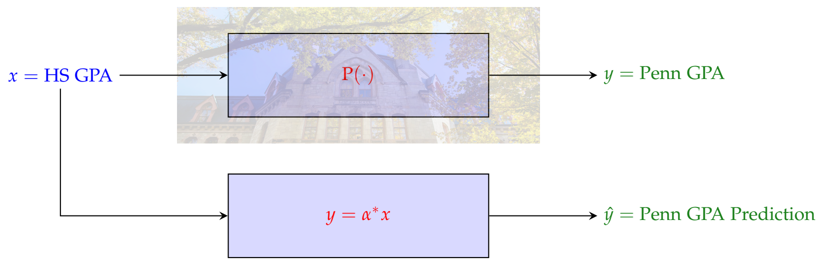

Figure 7. Penn grade point average prediction. We design an artificial intelligence that predicts Penn GPAs based on high school GPA and SAT scores. The AI mimics the same relationship observed in past students.

units, a prediction error of 0.094 is not very accurate. This is apparent in Figure 6 where the line of predicted Penn GPAs is a rough estimate of observed Penn GPAs.

3.3.3 Linear Regression

We consider now a more complete model in which the Penn GPA is deemed to depend on the high school GPA and the SAT score. We therefore define the input vector $x = [x_1; x_2] = [\text{HS GPA}; \text{SAT}]$

$$\text{Penn GPA} \;=\; y \;=\; P(x) \;=\; P \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} . \;=\; P \begin{bmatrix} \text{HS GPA} \\ \text{SAT} \end{bmatrix} . \tag{10}$$

To make Penn GPA predictions we postulate an input-output relationship of the form

$$\hat{y} \;=\; w_1 x_1 + w_2 x_2 \;=\; w^T x. \tag{11}$$

As is the case of (5), $\hat{y}$ in (11) is a *prediction* of the *true* Penn GPA y that will be available after the fact. The coefficient $w = [w_1; w_2]$ is to be determined with the goal of making predictions $\hat{y}$ close to to actual Penn GPA y.

Task 4 Define and compute the MMSE estimator coefficients w^*, that would extend the MMSE definition in (7). Show that this coefficient is given by the expression

$$w^* = \left[\sum_{i=1}^{N} x_i x_i^T \right]^{-1} \sum_{i=1}^{N} x_i y_i. \tag{12}$$

Figure 8. The University of Pennsylvania (Penn). Penn is a system that takes a student as an input and produces a graduate as an output.

Compute w^* for the data loaded in Task 1. Compute the RMSE of w^* and comment on the quality of the Penn GPA predictions. ■

4 Requirements

The AI of Section 3.3 is an indefensible strategy for making admission decisions to Penn. To talk about why this strategy is indefensible we need to talk about requirements. This is just a way of saying what is the goal of the AI system that we are designing. As it happens, we never made this goal explicit. However, implicit in the prediction of Penn GPAs is the fact that we intend to admit students with higher GPA potential. Thus, the specification (the requirement) of the AI system is the following:

> **(R1)** Admit the students that will attain the highest graduation GPA.

Making requirements explicit is important. If the AI makes decisions that are incompatible with our principles, it is not the model's fault, the data's fault, or the training process's fault. It is a problem of having misspecified requirements.

4.1 User Requirements

Requirement (R1) is a choice we made as engineers. This is not the same as the user requirement. The specification that people actually in charge of admission decisions would give.

This requirement is actually well known. Penn is a liberal arts institution. As such, our goal is to make students free. This is the literal meaning of liberal arts; the skill (art) of being free (liberal). Penn being Penn, the meaning of free is more concrete and was given to us by Benjamin Franklin[1]:

> "The Idea of what is *true Merit,* should also be often presented to Youth ..., as consisting in an *Inclination* join'd with an *Ability* to serve Mankind, one's Country, Friends and Family; which *Ability* is (with the Blessing of God) to be acquir'd or greatly encreas'd by *true Learning.*" [Emphasis mine]

Learn, so that you can develop the *inclination* and the *ability* to serve.

I did not attend Penn, but had I attended Penn, I expect the outcome would had been something like what is depicted in Figure 8. The student Alejandro would be transformed into the graduate Alejandro. The latter is more *libre*. More inclined and more able to serve mankind, his countries, his friends and his family. He would be happier for that.

Having this in mind, the following is a sensible requirement for the admission system:

> **(R2)** Admit the students that we can make the most free. Those that after attending Penn will be the most inclined and the most able to serve mankind and their countries, friends and families.

If you ever wondered why university admissions are so fraught, this is the reason. Penn does try to live up to its charter. We do believe in education as a service to our communities and we want to admit and teach students that can have the most positive impact in their communities.

This parenthetical comment aside, what is relevant here is the contrast between (R2) and (R1). Two remarks are warranted in this regard: (i) There is a lot of distance between the *user* requirement (R2) and the *engineering* requirement (R1). (ii) Requirement (R1) is partly motivated by the data that we have available. Indeed, Requirement (R1) reduces the beautiful complexity of an institution of higher education to a map between

[1] Benjamin Franklin, "Proposals Relating to the Education of Youth in Pennsylvania." October 1749.

performance indicators. These performance indicators are reducing the beautiful complexity of students and graduates to numbers and genders, both of which don't say much about their inclination and ability to serve mankind and their countries, friends and families. This coarse simplification of Penn is necessary, however, because the data that we have is limited.

Throughout, we will talk a lot about data, models and decisions but we won't talk much about requirements. This is because the focus of this course is on designing systems that satisfy given requirements. In actual engineering practice, requirements are flexible and it is important to think about how design considerations affect system requirements.

5 Report

Do not take much time to prepare a lab report. We do not want you to report your code and we don't want you to report your work. Just give us answers to questions we ask. Specifically give us the following:

Question	Report deliverable
Task 1	Plot of Penn GPA vs HS GPA
Task 1	Plot of Penn GPA vs SAT
Task 2	Derivation of (8)
Task 2	Value of α^*
Task 3	RMSE
Task 3	Comment on the quality of the Penn GPA predictions
Task 4	Derivation of (12)
Task 4	Value of w^*
Task 4	RMS
Task 4	Comment on the quality of the Penn GPA predictions

We will check that your answers are correct. If they are not, we will get back to you and ask you to correct them. As long as you submit responses, you get an A for the assignment. It counts for 10% of your lab grade.