Lecture 13: MATH 342W: Introduction to Data Science and Machine Learning

Sergio E. Garcia Tapia*

March 25, 2025 (last updated April 12, 2025)

Different Modeling Settings

We have explored a variety of modeling approaches in the context of different response spaces $\mathcal{Y}$. The table below shows some of these:

Response Space $\mathcal{Y}$	**Output of g belongs to**	**Name**	**Example Algorithm $\mathcal{A}$**
$\mathbb{R}$	$\mathbb{R}$	Regression	OLS
$\{0, 1\}$	$\{0, 1\}$	Binary Classification	Perceptron, SVM, KNN
$\{C_1, C_2, \ldots, C_L\}$ (Nominal)	$\{C_1, C_2, \ldots, C_l\}$ (Nominal)	(Multinomial) Classification	KNN
$\{0, 1\}$	$[0, 1]$	Probability Estimation	Logistic Regression

Beyond these, the following are discussed in 343:

Response Space $\mathcal{Y}$	**Output of g belongs to**	**Name**	**Example Algorithm $\mathcal{A}$**
$(0, \infty)$	$(0, \infty)$	Survival Modeling	Weibull Regression
$\{C_1, C_2, \ldots, C_L\}$ (Nominal)	$\boldsymbol{p}$ of dimension L	Probability Estimation	Multilogit Regression
$\{0, 1, 2, \ldots\}$	$\{0, 1, 2, \ldots\}$	Count Modeling	Poisson/Negative Binomial Regression

Even beyond the data science course sequence at QC:

Response Space $\mathcal{Y}$	**Output of g belongs to**	**Name**	**Example Algorithm $\mathcal{A}$**
$(0, 1)$	$(0, 1)$	Proportion Modeling	Beta Regression
$\{C_1, C_2, \ldots, C_L\}$	$\boldsymbol{p}$ of dimension L	Ordinal Modeling	Proportional Odds Regression

*Based on lectures of Dr. Adam Kapelner at Queens College. See also the course GitHub page.

Probability Estimation

Notice that in most models we surveyed, the response space $\mathcal{Y} = \{0, 1\}$ was identical to the output space of the prediction function g. In the table above, however, the entry for **logistic regression** is not like the others: $\mathcal{Y} = \{0, 1\}$ and g returns a number in the interval $[0, 1]$. The model function g returns an *estimate* of $P(Y = 1)$, the probability that the random variable $Y = 1$. In this setting, the null model g_0 is given by:

$$g_0 = \bar{y} = \text{the proportion of 1's in the } n \text{ observations in our data set } \mathbb{D}$$

Instead of considering $y = t(z_1, \ldots, z_t)$, we consider $Y \sim \text{Bernoulli}(t(z_1, \ldots, z_t))$, which is *not* random. As usual, we do not know t or the drivers $z_1, \ldots, z_t$, so we use the proxies $x_1, \ldots, x_p$. Thus y becomes random. Let

$$f_{pr} : \mathbb{R}^{p+1} \to (0, 1)$$

That is, f_{pr}, which is a function of the features, is a probability function. Realistically, we are never entirely sure if $y = 1$ or $y = 0$ (they are degenerate probabilities), so we omit them from the co-domain. We approximate Y as

$$Y \sim \text{Bernoulli}(f_{pr}(x_1, \ldots, x_p))$$

which differs based on $\boldsymbol{x}$. We can calculate the probability of seeing the historical data:

$$P(\mathbb{D}) = P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n \mid \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$$

Assuming that the observations are independent, we can simplify this joint probability function into a product of probabilities:

$$P(\mathbb{D}) = \prod_{i=1}^{n} P(Y_i = y_i \mid \boldsymbol{x}_i)$$

Recall that if $V \sim \text{Bernoulli}(p)$, then

$$P(V = v) = \begin{cases} p & \text{if } v = 1 \\ 1 - p & \text{if } v = 0 \\ 0 & \text{otherwise.} \end{cases}$$

Two other useful ways to write this PMF are:

$$P(V = v) = p^v (1 - p)^v \quad \text{and} \quad P(V = v) = pk + (1 - p)(1 - k),$$

where $v \in \{0, 1\}$ (the support of V). Then since Y is Bernoulli, we have

$$P(\mathbb{D}) = \prod_{i=1}^{n} f_{pr}(\boldsymbol{x}_i)^{y_i} (1 - f_{pr}(\boldsymbol{x}_i))^{1 - y_i}$$

Generalized Linear Models

Consider an algorithm $\mathcal{A}$ that finds a model by maximizing $P(\mathbb{D})$. Similar to before, f is not known; it could be arbitrarily complex. The set of functions that map from $\mathbb{R}^{p+1}$ to $(0, 1)$ is too large. Thus, let's consider a smaller candidate set $\mathcal{H}_{pr}$. As a first attempt, consider again the linear model (the set of hyperplanes):

$$\mathcal{H}_{pr} = \{\, \boldsymbol{w} \cdot \boldsymbol{x} \mid \boldsymbol{w} \in \mathbb{R}^{p+1} \}$$

Unfortunately, this does not work because the range of such functions is $\mathbb{R}$, and we only want values in the open interval $(0, 1)$. We would like to insist on a set of candidate functions that retain the linear model $\boldsymbol{w} \cdot \boldsymbol{x}$ for the following reasons:

(1) It is easily interpretable. For example, b is the change in y if x_1 changes by 1.

(2) Monotone in each feature, making it well-behaved.

In order to retain the simplicity of the linear model while still mapping into $(0, 1)$, we will *transform it*. We need a **link function**:

$$\phi : \mathbb{R} \to (0, 1)$$

This *links* $\mathbb{R}$ to a different space (i.e., $(0, 1)$). There are many such functions, but we will restrict our attention to strictly monotonically increasing ϕ. Then the set of candidate functions becomes:

$$H_{pr}(\phi) = \{\, \phi(\boldsymbol{w} \cdot \boldsymbol{x}) : \boldsymbol{w} \in \mathbb{R}^{p+1} \}$$

This is the space of **Generalized Linear Models (GLMs)**. In spite of our restriction to monotonically increasing functions, there are still infinitely-many link functions. Archetypal examples of link functions include all CDFs of continuous random variables whose support is all of $\mathbb{R}$.

Link Functions

The following are examples of link functions, listed in order of popularity:

(1) **Logistic**: The logistic (logit, or sigmoid) link function is precisely the PDF of the standard logistic random variable:

$$\phi(u) = \frac{e^u}{1 + e^u}$$

Often, we may see it in the following alternate form:

$$\phi(u) = \frac{e^u}{1 + e^u} \cdot \frac{e^{-u}}{e^{-u}} = \frac{1}{1 + e^{-u}}$$

On occasion we will also find it necessary to write:

$$1 - \phi(u) = \frac{1}{1 + e^u}$$

(2) **Probit**: The probit link function is the CDF of the standard normal random variable:

$$\phi(u) = \Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-v^2/2}\, dv$$
$$1 - \phi(u) = 1 - \Phi(u) = \Phi(-u)$$

(3) **Cloglog**: Standing for *complementary log-log*, it corresponds to the CDF of the standard Gumbel random variable:

$$\phi(u) = 1 - e^{-e^u} \iff 1 - \phi(u) = e^{-e^u}$$

Here is a bit of trivia: why is it called complementary log-log? We can see it by manipulating the link function:

$$\begin{aligned} \ln(1 - \phi(u)) &= -e^u \\ e^u &= -\ln(1 - \phi(u)) \\ u &= \underbrace{\ln(-\ln(\underbrace{1 - \phi(u)}_{\text{complement}}))}_{\text{log - log}} \end{aligned}$$

(4) **Cauchit**: Corresponds to

$$\phi(u) = \frac{1}{\pi} \arctan(u) + \frac{1}{2}$$

To see that this makes sense, recall that $\lim_{x \to \pm\infty} \arctan(u) = \pm\frac{\pi}{2}$. This is known as "exotic", and is useful when the distribution of x's are very variable, (for example, no expectation).

The most common are (1) logit and (2) probit.

Logistic Regression

Let ϕ be the logic link, and consider the set of candidate functions

$$\mathcal{H}_{pr}(\phi) = \left\{ \frac{1}{1 + e^{-\boldsymbol{w} \cdot \boldsymbol{x}}} \mid \boldsymbol{w} \in \mathbb{R}^{p+1} \right\}$$

We further approximate $Y \sim \text{Bernoulli}(h^*_{pr}(x_1, \ldots, x_p))$ (which adds misspecification error). Here, $h^*_{pr}(x_1, \ldots, x_p)$ is the best approximation of $f_{pr}(x_1, \ldots, x_p)$ in the set $\mathcal{H}_{pr}$. To choose an element of $\mathcal{H}_{pr}$ as our g, we need an algorithm $\mathcal{A}$, which for us will be:

$$\mathcal{A} : \boldsymbol{b} = \underset{\boldsymbol{w} \in \mathbb{R}^{p+1}}{\text{argmax}} \{ \underbrace{P(\mathbb{D})}_{\text{probability}} \} \tag{1}$$

This algorithm is called **logistic regression**, and the probability on $\mathbb{D}$ s given by

$$\begin{aligned} P(\mathbb{D}) &= \prod_{i=1}^{n} f_{pr}(\boldsymbol{x}_i)^{y_i} (1 - f_{pr}(\boldsymbol{x}_i))^{1-y_i} \\ &\approx \prod_{i=1}^{n} \phi(\boldsymbol{w} \cdot \boldsymbol{x}_i)^{y_i} (1 - \phi(\boldsymbol{w} \cdot \boldsymbol{x}_i))^{1-y_i} \\ &= \prod_{i=1}^{n} \left(\frac{1}{1 + e^{-\boldsymbol{w} \cdot \boldsymbol{x}_i}} \right)^{y_i} \left(\frac{1}{1 + e^{\boldsymbol{w} \cdot \boldsymbol{x}_i}} \right)^{1-y_i} \end{aligned}$$

When we discussed OLS, we used differentiation to find a vector $\boldsymbol{b}$ that minimized the SSE. Can we do something similar to maximize the probability, i.e., can we do

$$\frac{d}{d\boldsymbol{w}}\left(P(\mathbb{D})\right) \stackrel{\text{set}}{=} \mathbf{0}_{p+1}?$$

Unfortunately, there are no closed solutions to this as in the OLS case. To maximize the probability as specified in Equation 1, we need to use an approximation algorithm. In R, this might involve using `optimx`. The algorithm will yield a feature weight vector $\boldsymbol{b}$ that we can use to define:

$$\hat{p} := g_{pr}(\boldsymbol{x}) := \phi(\boldsymbol{b} \cdot \boldsymbol{x}) = \frac{1}{1 + e^{-\boldsymbol{b} \cdot \boldsymbol{x}}} \in (0, 1)$$

The vector $\boldsymbol{p}$ is distinct from $\hat{\boldsymbol{y}}$, because the latter would be a prediction (such as 0 or 1 in this the current response space of $\mathcal{Y} = \{0, 1\}$), whereas $\boldsymbol{p}$ is a probability, a real number in $(0, 1)$. The probability is our guess of $P(Y = 1 \mid \boldsymbol{x})$.

Since we worked hard to retain the linearity in our model (i.e., to retain the $\boldsymbol{w} \cdot \boldsymbol{x}$), let's see how we can use it. We isolate $\boldsymbol{b} \cdot \boldsymbol{x}$:

$$\begin{aligned}
\frac{1}{\hat{p}} &= 1 + e^{-\boldsymbol{b} \cdot \boldsymbol{x}} \iff \\
e^{-\boldsymbol{b} \cdot \boldsymbol{x}} &= \frac{1}{\hat{p}} - 1 = \frac{1 - \hat{p}}{\hat{p}} \iff \\
-\boldsymbol{b} \cdot \boldsymbol{x} &= \ln\left(\frac{1 - \hat{p}}{\hat{p}}\right) \iff \\
\boldsymbol{b} \cdot \boldsymbol{x} &= \ln\left(\frac{\hat{p}}{1 - \hat{p}}\right)
\end{aligned} \tag{2}$$

The quantity $\frac{\hat{p}}{1-\hat{p}} \in (0, \infty)$ is called the **odds of** $Y = 1$, and $\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) \in \mathbb{R}$ is called the **log-odds of** $Y = 1$. Both of these have a one-to-one relationship with $P(Y = 1)$. We say that *our model is linear in log-odds.*

What is the relationship between $Y = 1$ and $P(Y = 1)$? If log-odds is 0, then the probability is $\frac{1}{1+e^{-0}} = \frac{1}{2}$. The following table shows a worthwhile subset of values to be familiar with:

Log-odds	$P(Y = 1 \mid \mathbf{x})$
0	$\frac{1}{2}$
-1	$0.27 \approx \frac{1}{4}$
1	$0.73 \approx \frac{3}{4}$
-2	$0.12 \approx \frac{1}{8}$
2	$0.88 \approx \frac{7}{8}$
-3	0.05
3	0.95
-4	0.02
4	0.98
$-\infty$	0
∞	1

See also Figure 1. Note that depending on the current value of the log-odds, an increase by 1 may or may not lead to a substantial increase in probability. For example:

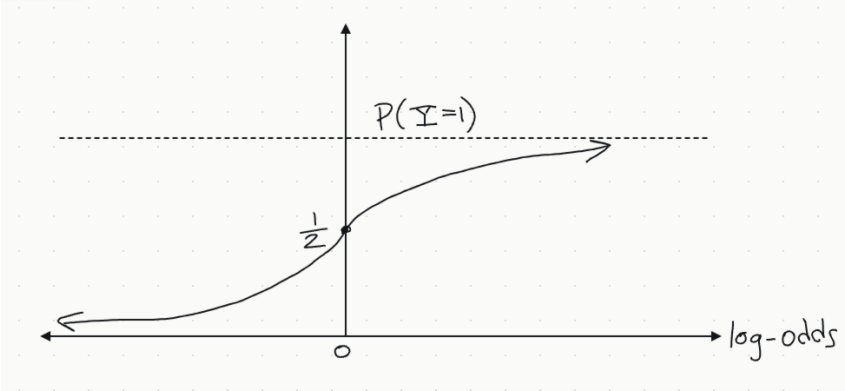

Figure 1: A plot of log-odds versus probability that $Y = 1$.

- When the log-odds increases from 0 to 1, the probability increases from 0.5 to 0.73.
- When the log-odds increases from 3 to 4, the probability increases from 0.95 to 0.98 (a very small change).

Performance Metrics

Previously we used $RMSE$ and R^2 to assess the predictions of our model. For logistic regression, we consider the following **scoring rules**:

(1) **Brier Score**: Let

$$s_i := -(y_i - \hat{p}_i)^2$$

Then the *Brier score* is defined by:

$$\bar{s} := \frac{1}{n}\sum_{i=1}^{n} s_i$$

Note $\bar{s} \leq 0$, since all the $s_i \leq 0$. There are two special cases to consider:

- The *best* Brier score is 0, when $\hat{p}_i = y_i$ (that is, the probability is 1 when the response is 1, and the probability is 0 when the response is 0). Then $s_i = 0$ for all i, and hence $\bar{s} = 0$.
- The *worst* Brier score is -1. It occurs when $\hat{p}_i = 1 - y_i$ for all i, so that $s_i = -1$ for all i, and hence $\bar{s} = -1$.
- Yet another important case is when $\hat{p}_i = \frac{1}{2}$; the model is confused and it becomes a coin toss. In this case,

$$\bar{s} = \sum_{i=1}^{n}\left(y_i - \frac{1}{2}\right)^2 = -\frac{1}{4}$$

In the homework, you will show that the null model g_0 beats this Brier score.

(2) **Log Scoring**: This time, we define

$$s_i := y_i \ln(\hat{p}_i) + (1 - y_i) \ln(1 - \hat{p}_i)$$

which also satisfies $s_i \leq 0$, since $0 < \hat{p}_i < 1$ and hence both logarithmic expression evaluate to negative. Note that this comes from the Bernoulli distribution.

Then the *log scoring rule* is given by

$$\bar{s} := \frac{1}{n} \sum_{i=1}^{n} s_i$$

and again we have $\bar{s} \leq 0$. Here, 0 is best log score value. Meanwhile, if $\hat{p}_i = \frac{1}{2}$ for all i, we have

$$\bar{s} = \frac{1}{n} \sum_{i=1}^{n} \ln\left(\frac{1}{2}\right) = \ln\left(\frac{1}{2}\right)$$