Week 9:
Logistic Regression

CMPS 320 : Machine Learning

An Overview of Classification

- The linear regression model assumes that the response variable Y is quantitative.
- In many situations, the response variable is instead qualitative.
- Examples of qualitative variables are sex, marital status, political affiliation etc
- We will study approaches for predicting qualitative responses, a process that is known as **classification**.
- Predicting a qualitative response for an observation can be referred to as **classifying** that observation, since it involves assigning the observation to a category, or class.

An Overview of Classification (cont.)

- In the classification setting we have a set of training observations $(x_1, y_1), \cdots, (x_n, y_n)$ that we can use to build a classifier.
- We want our classifier to perform well not only on the training data, but also on test observations that were not used to train the classifier.

Why Not Linear Regression?

- Linear regression works on quantitative responses.
- Suppose that we are trying to predict the medical condition of a patient in the emergency room on the basis of her symptoms.
- In this example, there are three possible diagnoses: stroke, drug overdose, and epileptic seizure.
- We can encode these values as a quantitative response variable, Y, as follows:

$$Y = \begin{cases} 1 & \text{if stroke} \\ 2 & \text{if drug overdose} \\ 3 & \text{if epileptic seizure} \end{cases}$$

Why Not Linear Regression?

- Using this coding, least squares could be used to fit a linear regression model to predict Y on the basis of a set of predictors $X_1, \cdots, X_p$.
- Two issues arise:
  - This coding implies an ordering on the outcomes, putting drug overdose in between stroke and epileptic seizure.
  - The coding also insist that the difference between stroke and drug overdose is the same as the difference between drug overdose and epileptic seizure.
- In practice there is no particular reason that this needs to be the case.
- There is no natural way to convert a qualitative response variable with more than two levels into a quantitative response that is ready for linear regression.

Logistic Regression

- Consider a credit default data set, where the response default falls into one of two categories, Yes or No.
- Rather than modeling this response Y directly, **logistic regression** models the probability that Y belongs to a particular category.
- Logistic regression models the probability of default:
  - For example, the probability of default given balance can be written as:

$$Pr(\text{default} = \text{Yes}|\text{balance})$$

  - The values of $Pr(\text{default} = \text{Yes}|\text{balance})$ is abbreviated $p(\text{balance})$ will range between 0 and 1.
  - Then for any given value of balance, a prediction can be made for default.

Logistic Regression (cont.)

- For example, one might predict **default = Yes** for any individual for whom $p(\text{balance}) > 0.5$.
- Alternatively, if a company wishes to be conservative in predicting individuals who are at risk for default, then they may choose to use a lower threshold, such as $p(\text{balance}) > 0.1$.

Logistic Regression (cont.)

- How should we model the relationship between $p(X) = Pr(Y = 1|X)$ and X?
  - We must model $p(X)$ using a function that gives outputs between 0 and 1 for all values of X.
- Several functions meet this description for example: the logistic function.
- The logistic function - noted $\sigma(\cdot)$ – is a sigmoid function (i.e., S-shaped) that outputs a number between 0 and 1.

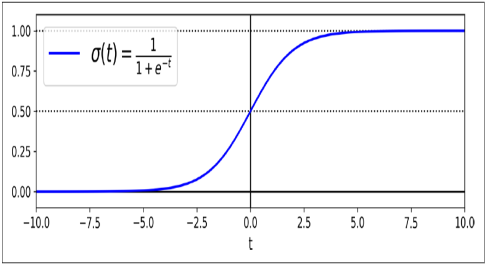

Figure 4-21. Logistic function

Logistic Regression (cont.)

- In logistic regression, we use a logistic function:

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} \tag{1}$$

  where e is the natural logarithm base
- To fit the logistic model we use a method called **maximum likelihood**.
- The logistic function will always produce an S-shaped curve and regardless of the value of X, we will obtain a sensible prediction.

Transforming Logistic Regression

- Equation (1) can be transformed as:

p(happens)

$$e^{\beta_0+\beta_1 \mathrm{X}} = \frac{p(\mathrm{X})}{1-p(\mathrm{X})}$$ **“odds”**

p(doesn’t)

$$\log(e^{\beta_0+\beta_1 \mathrm{X}}) = \log\left(\frac{p(\mathrm{X})}{1-p(\mathrm{X})}\right)$$ **“log odds”** a.k.a. logit

$$\boxed{\beta_0+\beta_1\mathbf{x}}$$

- The quantity $p(X)/[1-p(X)]$ is called the odds.
    - The odds takes value between 0 and ∞.
    - Values of the odds close to 0 indicate very low probabilities of default and
    - Values of the odds close to ∞ indicate very high probabilities of default.
- **Interpretation of** β_1: Increasing X by one unit changes the log odds by β_1 or it multiplies the odds by e^{β_1}

Estimating Logistic Regression coefficients with maximum likelihood

- The coefficients β_0 and β_1 are unknown, and must be estimated based on the available training data.
- In logistic regression, we want coefficients that yield:
  - values close to 1 (high probability) for observations in the class
  - values close to 0 (low probability) for observations not in the class
- We can formalize this intuition mathematically using a likelihood function:

$$\boxed{\prod_{i:y_i=1} p(x_i)} \times \boxed{\prod_{j:y_j=1} \left(1 - p(x_j)\right)}$$

- The goal is to estimate coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ that maximize this function.

Making Predictions

- Making predictions is pretty straightforward after estimating the coefficients:
- We use the equation:

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X}}$$

where e is the natural logarithm base.

Multiple Logistic Regression

- We now consider the problem of predicting a binary response using multiple predictors:

$$log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1(x)$$

$$log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1(x_1) + \cdots + \beta_k(x_k)$$

- Where $X = (X_1, \ldots, X_p)$ are p predictors.
- The above equation can be rewritten as:

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}} \tag{2}$$

- We use maximum likelihood method to estimate the coefficients $\beta_0, \beta_1, \cdots, \beta_p$