Introduction to classification

- For linear regression, the response variable Y is quantitative.
- Classification problems deal with the situation when the response variable is qualitative or categorical.
- Predicting a qualitative response for an observation is called classifying that observation.
- Logisstic regression and linear disrciminant analysis first predict the probability of each of the categories of a qualitative variables.

Classification

- Classification setting: training observations $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$. y_is are categorical.
- Examples: Digit recognition: Inputs are images of digits (could be handwritten or photographed), response variables are the labels of those digits, 0 to 9. An automated classification algorithm can build a classifider from training data to classify an unknown image.
- Training data: DNA sequence of a number of patients with or without a given disease, a biologist wants to figure out which DNA mutations are disease causing and which are not.
- In this course, we mostly deal with binary classfication (or two-class).

Classification

- Binary classification: two class Success and Failure(1 vs. -1) (or 1 vs. 0).
- Why linear regression can't be used in this case? Some estimates could be way out of the range [0,1] (or [-1,1]).
- Logistic regression models the probability that response Y belongs to a particular class.
- Given input data X, denote $Pr(Y = 1|X)$=p(X). If $p(X) > 0.5$ (or some threshold), predicts response Y=1 (Success). Otherwise predicts Y=0 (or -1) (Failure).

Logistic model

- Need to model p(X) using a function that gives outputs between 0 and 1 for all values of X by using a logistic function.

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

  This is equivalent to:

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

  $\frac{p(X)}{1-p(X)}$ is called the odds (chance of winning / chance of losing). It can take values between 0 and ∞.
- Interesting property: p(-X)=1 - p(X).

Logistic function

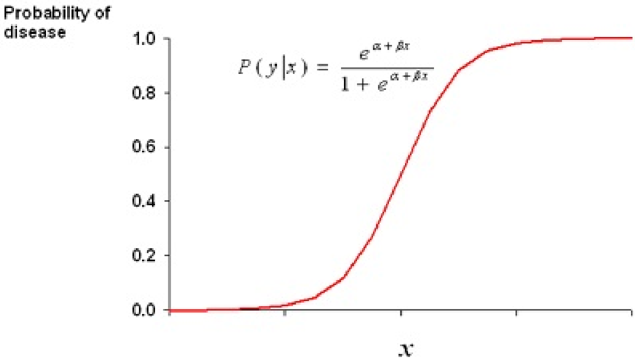

Figure: onlinecourses.science.psu.edu

Logistic model

- Take the log of both side we obtain:

$$log(\frac{p(X)}{1-p(X)}) = \beta_0 + \beta_1 X = \beta^T X.$$

- In logistic models, instead of directly modeling the response Y as a linear function of predictors X, we model the log odds instead.

Maximum likelihood estimation

- For linear regression, we used least squares approach to estimate linear regression coefficients.
- For logistic regression, we don't have information on p(X). But we are trying to estimate p(X).
- A general method of maximum likelihood: estimate coefficient β so that the model fits the data as "good" as possible.
- Given a data point (x, y), y can be considered a random sample from an unknown distribution. In our case, a Bernoulli trial.

Bernoulli trial

- Bernoulli trial is a random experiment with exactly two possible outcomes: "success" or "failure", in which the probability of success if the same every time.
- Probability of success usually denoted by : p.
- Bernoulli distribution is a distribution of a random variable which takes the value of 1 with probability p, 0 with probability 1-p.
- Pdf function: $f(outcome) = p^{outcome}(1-p)^{1-outcome}$, outcome $=$0,1.
- Example: Flip a coin.

Maximum likelihood estimation

- Given data points $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ generated from some distribution with density function and parameter θ: $f(x_1, x_2, \cdots, x_n, \theta)$
- The likelihood function is a function in terms of parameter θ: $\mathcal{L}(\theta) = f(x_1, x_2, \cdots, x_n|\theta)$.
- $\mathcal{L}(\theta)$ is the probability of observing the given data as a function of θ.
- The maximum likelihood estimator of θ is the value of θ that maximizes θ.

Maximum likelihood estimation

- If the observations are identically independent distributed (iid), the likelihood is simplified to: $\mathcal{L} = \prod_{i=1}^{n} f(x_i|\theta)$.
- Maximize a function of product is quite difficult so we use a log transform and obtain the problem

$$\theta^* = \max \sum_{i=1}^{n} log(f(x_i|\theta)).$$

- In our logistic model, each observation (x, y) has probability of success p(x) and failure probability 1-p(x). Therefor the likelihood function that we need to maximize is:

$$\mathcal{L}(\beta) = \sum_{i:y_i=Success} log(p(x_i)) + \sum_{i:y_i=Failure} log(1 - p(x_i))$$

Maximum likelihood esstimation

- We have already denoted:

$$P(y = Success|\beta, x) = \frac{e^{\beta^T x}}{1 + e^{\beta^T x}}$$
$$P(y = Failure|\beta, x) = \frac{1}{1 + e^{\beta^T x}}$$

- Denote y=1 for Success and y=-1 for Failure, we can show that:

$$P(y = \pm 1|\beta, x) = \frac{1}{1 + e^{-y\beta^T x}}$$

Maximum likelihood

- When $y = 1$ we can see that:

$$\frac{1}{1+e^{-y\beta^T x}} = \frac{1}{1+e^{-\beta^T x}} = \frac{e^{\beta^T x}}{1+e^{\beta^T x}} = P(y = Success|\beta, x).$$

- Similarly When $y = -1$ we can see that:

$$\frac{1}{1+e^{-y\beta^T x}} = \frac{1}{1+e^{\beta^T x}} = P(y = Failure|\beta, x).$$

MLE

- Now we can write an explitcit formula for the likelihood function:

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} log(\frac{1}{1+e^{-y_i\beta^T x_i}})$$
$$\beta^* = \max \mathcal{L}(\beta).$$

$$\beta^* = \max \sum_{i=1}^{n} [log 1 - log(1+e^{-y_i\beta^T x_i})]$$

- Notice that if x^* is the maximizer of f(x) then x^* is also the minimizer of the function -f(x).
-

$$\beta^* = \min_{\beta} \sum_{i=1}^{n} log(1+e^{-y_i\beta^T x_i})$$

MLE solution

- The estimator of logistic regression β^* is the minimizer of the following problem that minimize the negative log-likelihood:

$$\min_{\beta} f(\beta) = \min_{\beta} \sum_{i=1}^{n} log(1 + e^{-y_i \beta^T x_i})$$

- Unlike linear regression, it is not possible to have an explicit solution to this problem.
- We need to use a computer procedure to calculate the solution.

Solution to logistic regression

- Gradient descent method can be used to find the optimal solution of the problem above.
- Consider a small example with 2 observations: (x_1, y_1) and (x_2, y_2): Probability of success:

$$p(y_1 = 1|x1) = \frac{1}{1 + e^{-(\beta_1 x_{11} + \beta_2 x_{12})}}$$
$$p(y_2 = 1|x2) = \frac{1}{1 + e^{-(\beta_1 x_{21} + \beta_2 x_{22})}}$$

$$f(\beta) = log(1 + e^{-y_1(\beta_1 x_{11} + \beta_2 x_{12})}) + log(1 + e^{-y_2(\beta_1 x_{21} + \beta_2 x_{22})})$$

- Take derivative in terms of β_1:

$$\frac{\partial f}{\beta_1} = \frac{-y_1 x_{11}}{1 + e^{-y_1(\beta_1 x_{11} + \beta_2 x_{12})}} + \frac{-y_2 x_{21}}{1 + e^{-y_2(\beta_1 x_{21} + \beta_2 x_{22})}}$$

$$\frac{\partial f}{\beta_2} = \frac{-y_1 x_{12}}{1 + e^{-y_1(\beta_1 x_{11} + \beta_2 x_{12})}} + \frac{-y_2 x_{22}}{1 + e^{-y_2(\beta_1 x_{21} + \beta_2 x_{22})}}$$

- Gradient of the function evaluated at a point β

$$\nabla f(\beta) = \sum_{i=1}^{n} \frac{y_i x_i^T}{1 + e^{-y_i \beta^T x_i}}$$

- With this gradient, we can use gradient descent method to find solution for logistic regression.
- Matlab demonstration

How to make prediction

-

MLE

- Earlier, we mentioned if $P(y = Success|x) \geq P(y = Failure|x)$ then we classify the observation as Success. This means:

$$\frac{e^{\beta^T x}}{1 + e^{\beta^T x}} \geq \frac{1}{1 + e^{\beta^T x}}$$
$$exp(\beta^T x) \geq 1$$
$$\beta^T x \geq 0.$$

- Interpretation: The "line" $\beta^T x$ (actually its called hyperplane) seperate the two Success and Failure classes. Points are classified as to which side of the "line" they are located.

Classfication rule

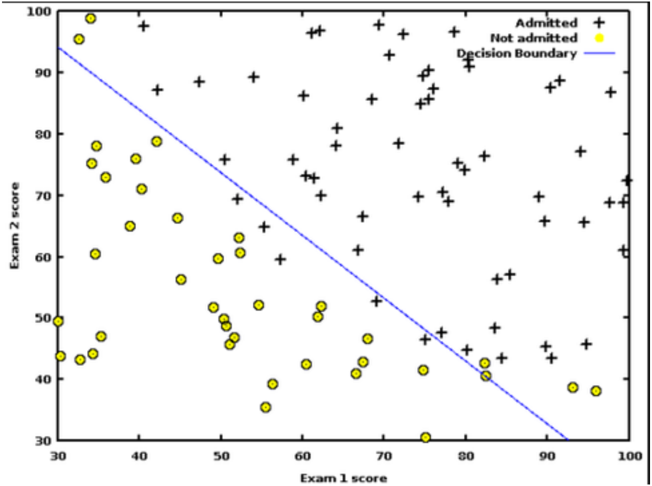

Figure: https://qph.is.quoracdn.net