5 Classification

Classification refers to the case when Y (response) is **categorical** or **qualitative**.

Ex
- Will a person default on their mortgage?
- " " " have heart disease as they age?
- Is an email spam?
- Will a certain ad make it more likely for someone to buy a product?

Note As before, still have predictors, and the response can be either binary (0/1) or multinomial.

Methods
- Logistic regression
- Discriminant analysis
- k-nearest neighbors
- support vector machines

Remark If $Y \in \{0,1\}$, don't want to do usual regression

$$\underset{\sim}{Y} = X\underset{\sim}{\beta} + \varepsilon$$

[why?] $X\hat{\underset{\sim}{\beta}}$ never 0 or 1

Alternative Model $P(Y=0)$ or $P(Y=1)$

The classifier / classification rule is a rule that assigns some probabilities of 1 and others to 0.

Ex: If $P(Y=1) \geq \frac{1}{2} \Rightarrow \hat{Y} = 1$

5.1 Logistic Regression

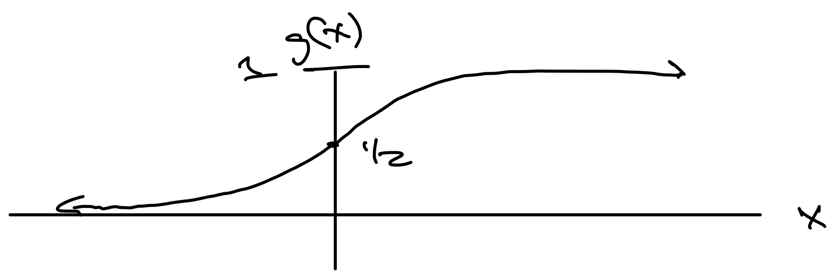
Setup $Y \in \{0,1\}$ is binary response & X is a single predictor.

Goal Model $P(Y=1 \mid X) := p(X)$

$p(X) = \beta_0 + \beta_1 X$? Bad idea b/c $\beta_0 + \beta_1 X \in \mathbb{R}$

DEF The logistic function is

$$g(x) = \frac{e^x}{1+e^x} \in (0,1)$$



logistic regression uses

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \quad \left[= \mathbb{P}(Y=1 \mid x) \right] = \text{function of } x$$

[Aside: other functions could be used, e.g. probit regression
$p(x) = \Phi(\beta_0+\beta_1 x)$ where Φ is the cdf of $N(0,1)$]

$\boxed{\text{Note}}$ The logit transform is the inverse of the logistic fcn,

$$f(p) = \log\left(\frac{p}{1-p}\right)$$

So for logistic regression $\Rightarrow$

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

$\Rightarrow$ we assume the log odds ratio is linear in x.

$$\frac{p}{1-p} = \text{odds ratio}$$

$p = \frac{1}{2} \Rightarrow$ odds $= 1 =$ even

$p = 0.2 \Rightarrow$ odds $= \frac{1}{4}$

Assumption is log odds ratio increases by β_1 for a unit increase in x.

Note The sign of β_1 tells us the effect of x on $p(Y=1)$

$\beta_1 > 0 \Rightarrow$ prob grows with x

$\beta_1 < 0 \Rightarrow$ " shrinks " "

Estimation In setup $Y \in \{0,1\} \Rightarrow Y \sim \text{Bernoulli}(p(x))$

Suppose we have indep data $(x_1, y_1), \ldots, (x_n, y_n)$, then the likelihood fcn for $\underline{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$

$$f(\underline{y}) = \prod_{i=1}^{n} p(x_i)^{y_i} (1 - p(x_i))^{1-y_i} \qquad [= f(\beta_0, \beta_1)]$$

Estimate β_0 + β_1 by maximizing $f(\underline{y})$ (AKA maximum likelihood estimation). Closed form MLEs are not available, so maximize f numerically.

$\boxed{\text{Sanity check}}$ No features: $p(\cancel{x}) = p = \dfrac{e^{\beta_0}}{1+e^{\beta_0}}$

Then

$$f(\underline{x}) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i} = p^{n\bar{Y}}(1-p)^{n-n\bar{Y}}$$

take deriv wrt p

$$\frac{\partial}{\partial p}\log f = \frac{n\bar{Y}}{p} - \frac{n-n\bar{Y}}{1-p} \overset{\text{set}}{=} 0$$

$\Rightarrow \hat{p} = \bar{Y}$ = proportion of 1s in data

$\hat{\beta}_0 = \log\left(\dfrac{\bar{Y}}{1-\bar{Y}}\right)$ = ~~log~~ odds ratio of empirical P(success) to P(failure)

Multiple logistic regression follows:

$$p(Y=1 \mid X_1, \ldots, X_p) = p(X_1, \ldots, X_p) = p(\underline{X})$$

$$= \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}} = \frac{e^{\beta_0 + \underline{\beta}^T \underline{X}}}{1 + e^{\beta_0 + \underline{\beta}^T \underline{X}}}$$

where $\underline{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$ no intercept

$$\iff \log\left(\frac{p(\underline{X})}{1 - p(\underline{X})}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

$\boxed{\text{Note}}$ Can use z-statistics $\dfrac{\hat{\beta}_j}{\widehat{SE}(\hat{\beta}_j)}$ to do hypothesis testing.

Can construct CIs.

$\boxed{\text{prediction}}$ For new set of features $\underline{x} = (x_1, \ldots, x_p)^T$

our predictor for $p(\underline{x})$ is

$$\hat{p}(\underline{x}) = \hat{P}(Y=1 \mid \underline{x}) = \frac{e^{\hat{\beta}_0 + \hat{\underline{\beta}}^T \underline{x}}}{1 + e^{\hat{\beta}_0 + \hat{\underline{\beta}}^T \underline{x}}} \in (0,1)$$

Need a way to convert $\hat{p}$ to $\hat{y} \in \{0,1\}$, the usual <u>classification rule</u> is

$$\hat{y} = \begin{cases} 1 & \hat{p}(\underline{x}) > \frac{1}{2} \\ 0 & \hat{p}(\underline{x}) < \frac{1}{2} \end{cases}$$

with randomization at $\hat{p} = \frac{1}{2}$.

$\hat{y} = 1$ if

$$\frac{1}{2} < \hat{p} \iff \frac{1}{2} < \frac{e^{\hat{\beta}_0 + \hat{\beta}^T \underline{x}}}{1 + e^{\hat{\beta}_0 + \hat{\beta}^T \underline{x}}}$$

$$\iff \frac{1}{2} < \frac{1}{2} e^{\hat{\beta}_0 + \hat{\beta}^T \underline{x}}$$

$$\iff 0 < \hat{\beta}_0 + \hat{\beta}^T \underline{x}$$

The classification rule is:

$$\hat{y} = \begin{cases} 1 & \hat{\beta}_0 + \hat{\beta}^T \underline{x} > 0 \\ 0 & \hat{\beta}_0 + \hat{\beta}^T \underline{x} < 0 \end{cases}$$

The line $\hat{\beta}_0 + \hat{\beta}^T \underline{x} = 0$ is the <u>decision boundary</u>.