Week 3 - **Logistic Regression**

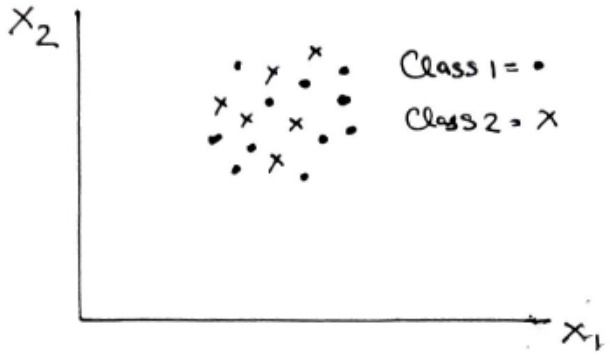

Figure 0.2: Classification

Problem setup:

1. n = Number of observations
2. p = Number of predictor variables (excluding the constant 1)
3. $y \in \{0, 1\}$ = Dependent variable (binary outcome, yes or no, qualitative)
4. $x_1, \ldots, x_p$ = Independent variables (predictors)

This problem can be viewed as a classification problem where the problem is to predict which class (0 or 1) an observation belongs to or alternatively as a regression problem where the probability that the output is of a particular class (a number in the range $[0, 1]$) is predicted in terms of the independent variables.

We are interested in estimating the linear model:

$$y^* \quad = \quad \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon.$$

However in this case, we do not observe y^* (a latent variable), but rather we observe a variable y which is defined as follows:

$$y = 1 \text{ if } y^* \geq 0, \text{ and } y = 0 \text{ otherwise.}$$

Key ideas:

1. The problem is to estimate $P(y = 1)$ and $P(y = 0)$ in terms of the predictor variables $\{x_1, \ldots, x_p\}$. Using a linear regression model is not suitable since the probability must lie between 0 and 1.

   The logistic function (S-shaped) provides a nice formulation to capture this:

   $$P(y = 1) \quad = \quad \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p}}.$$

   This number is always between 0 and 1, irrespective of the value of the x variables.

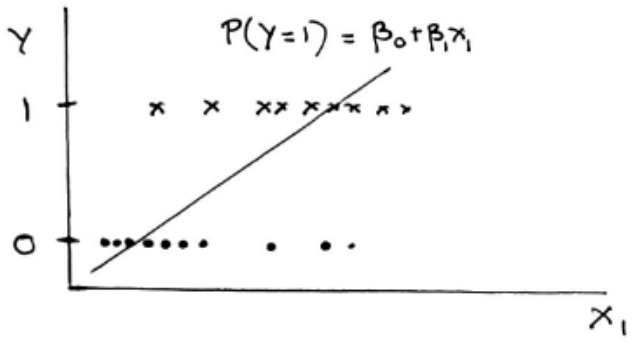

Figure 0.3: Linear regression

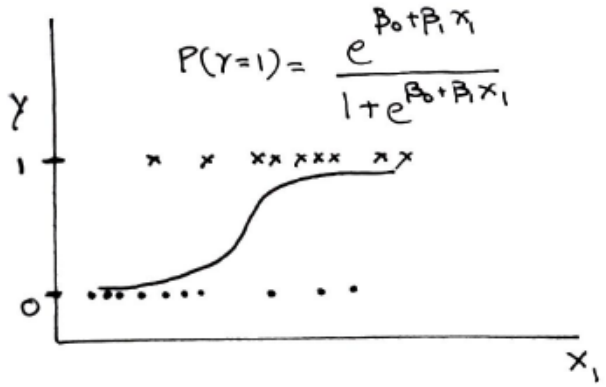

Figure 0.4: Logistic regression

2. Odds is defined as:

$$\begin{aligned} \text{Odds} &= \frac{P(y=1|x)}{P(y=0|x)} \\ &= e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p}. \end{aligned}$$

Odds > 1 if $y = 1$ is more likely and Odds < 1 if $y = 0$ is more likely, given a particular x.

The logit or log-odds is defined as:

$$\text{Log(Odds)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p,$$

and is linear in the beta coefficients. Note that a positive β_j coefficient indicates that $P(y = 1)$ increases, if x_j increase. Similarly, a negative β_j coefficient indicates that $P(y = 1)$ decreases, if x_j increases. However in contrast to linear regression, increasing the x_j variable by 1 unit (keeping all of the other x_k values fixed), changes the log-odds by β_j or multiplies the odds by e^{β_j}.

3. Define the coefficients to be estimated as:

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}.$$

To estimate the coefficients, we maximize the likelihood function which is defined as:

$$\text{Likelihood}(\beta) = L(\beta) = \prod_{i=1}^{n} P(y = 1|x = x_i; \beta)^{y_i} P(y = 0|x = x_i; \beta)^{1-y_i}.$$

The maximum likelihood problem is defined as:

$$\max_{\beta} \prod_{i=1}^{n} P(y = 1|x = x_i; \beta)^{y_i} P(y = 0|x = x_i; \beta)^{1-y_i},$$

where the objective function is the likelihood of the observations, assuming that each observation is independent of the other. The beta coefficients are estimated to maximize the probability of observing the actual responses under the model for each observation $i = 1, \ldots, n$. This problem is equivalently solved by the maximum log-likelihood problem:

$$\max_{\beta} \sum_{i=1}^{n} y_i \log(P(y = 1|x = x_i)) + (1 - y_i) \log(P(y = 0|x = x_i)),$$

where the log-likelihood is defined as:

$$\text{Log-Likelihood}(\beta) = LL(\beta) = \sum_{i=1}^{n} y_i \log(P(y = 1|x = x_i; \beta)) + (1 - y_i) \log(P(y = 0|x = x_i; \beta)).$$

Equivalently, this can be rewritten as:

$$\max_{\beta} \sum_{i=1}^{n} y_i \left(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}\right) - \sum_{i=1}^{n} \log\left(1 + e^{\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}}\right),$$

This objective function is concave in the beta variables and we are solving a maximization problem over these variables. In this case, the local optimum is the global optimum and the problem is efficiently solvable. However, we need to use an iterative algorithm to find the optimal solution in this case.

Algorithm: Define for $i = 1, \ldots, n$:

$$x_i = \begin{bmatrix} 1 \\ x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{bmatrix}.$$

The optimality condition is given by the system of equations:

$$\nabla h(\beta) := \sum_{i=1}^{n} \left(y_i - p_i(\beta' x_i)\right) x_i = 0,$$

where $p_i(\beta' x_i) = \dfrac{e^{\beta' x_i}}{1 + e^{\beta' x_i}}$. Let y denote a column vector of size n with entries y_i and X as the input data matrix of size $n \times (1 + p)$ with rows denoting the observations and columns denoting the intercept and the predictors. Applying Newton's method to solve this problem gives the iterates:

$$\beta^{(k+1)} = \beta^{(k)} + (\nabla^2 h(\beta^{(k)}))^{-1} \nabla h(\beta^{(k)}).$$

The gradients and the Hessian matrix are computed as:

$$\nabla h(\beta) = X'(y - p(\beta, X))$$

$$\nabla^2 h(\beta) = -X' W(\beta, X) X$$

where $p(\beta, X)$ is a vector with entries as $p_i(\beta' x_i)$ and $W(\beta, X)$ is a diagonal matrix with entries $p_i(\beta' x_i)(1 - p_i(\beta' x_i))$.

Quality of fit:

1. Deviance is a measure of the fit of the generalized linear model (of which logistic regression is an example) where higher numbers indicates a worse fit.

   Null deviance measures how well the response variable is predicted by a model that only includes the intercept.

   Residual deviance measures how well the response variable is predicted by the intercept and the p predictor variables.

   A significant decrease in the value from the null to residual deviance indicates that the predictor variables are useful in explaining the response variable. For logistic regression problems, suppose we estimate these coefficients as $\hat{\beta}_0$ and $\hat{\beta}$, then:

   $$\text{Null deviance } = -2LL(\hat{\beta}_0),$$

   where $\hat{\beta}_0$ is the estimated coefficient when only the intercept is fit and

   $$\text{Residual deviance } = -2LL(\hat{\beta}).$$

   We can verify that when only $\hat{\beta}_0$ is fit in the maximum likelihood estimation, it is chosen that the estimated fraction of 1's from the model is exactly equal to the observed fraction of 1's in the training set.

   The Akaike information criterion (AIC) is based on deviance but penalizes for making the model more complicated (similar to adjusted R-squared). However the AIC does not have a benchmark range like R-squared which lies in the inteval $[0, 1]$. Smaller the AIC, the better is the model. AIC is defined as:

   $$AIC = -2LL(\hat{\beta}) + 2(p + 1),$$

   where the p coefficients and the intercept coefficient are estimated.

2. Confusion matrix: The confusion matrix is a table that allows to visualize the performance of a classification algorithm as follows.

   Suppose we use the following rule to classify or predict an output:

   (a) Choose a threshold t

   (b) For any observation i with predictor variables x_i and estimated coefficients $\hat{\beta}$:
   If $P(y = 1 | x_i; \hat{\beta}) \geq t$, then predict 1, else if $P(y = 1 | x_i; \hat{\beta}) < t$, predict 0

	Actual = 0	Actual = 1
Predict = 0	True Negative (TN)	False Negative (FN)
Predict = 1	False Positive (FP)	True Positive (TP)

Table 0.1: Confusion matrix

This results in the confusion matrix as follows:

We can define the following quantities:

$$\begin{aligned}
\text{False positive rate:} \quad & FPR = && \frac{FP}{FP+TN} && \text{(Type I error)},\\
\text{True negative rate:} \quad & TNR = && \frac{TN}{FP+TN} && \text{(Specificity)},\\
\text{True positive rate:} \quad & TPR = && \frac{TP}{TP+FN} && \text{(Sensitivity)},\\
\text{False negative rate:} \quad & FNR = && \frac{FN}{TP+FN} && \text{(Type II error)},\\
\text{Overall Accuracy} \quad & \mathit{Accuracy} = && \frac{TP+TN}{FP+TN+TP+FN}. &&
\end{aligned}$$

Note that by definition, $FPR+TNR=1$ and $TPR+FNR=1$. By varying the threshold, we change the entries in the confusion matrix and the corresponding FPR, TPR, FNR, TNR.

3. Receiver operating characteristic (ROC) curve: Rather than computing the TPR and FPR for just a fixed threshold t, the ROC curve plots $TPR(t)$ and $FPR(t)$ as a function of t

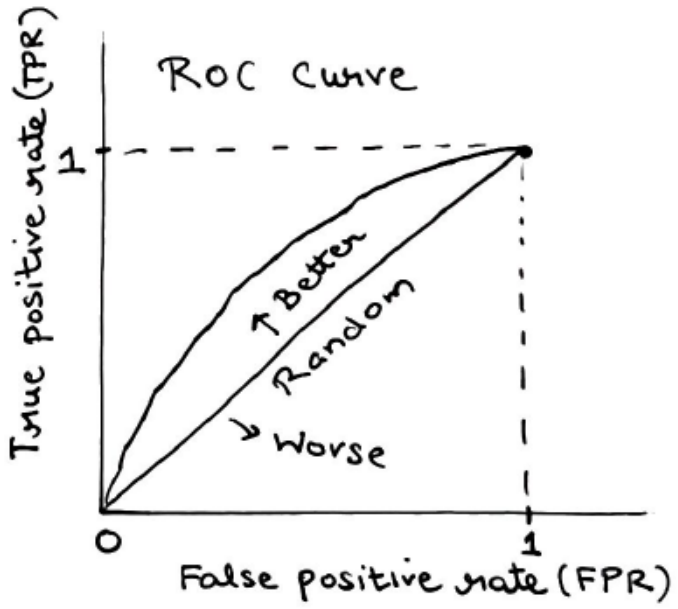

Figure 0.5: ROC Curve

   (a) Setting $t = 0$: All predictions are $y = 1$. Then $FPR = 1$ and $TPR = 1$.

   (b) Setting $t = 1$: All predictions are $y = 0$. Then $FPR = 0$ and $TPR = 0$.

   (c) If a model is performing at the level of chance (randomly guessing by flipping a coin), then we can achieve a point along the diagonal with $FPR = TPR$.

   (d) A system that perfectly separates $y = 1$ from $y = 0$ has a ROC curve with $FPR = 0$ and $TPR = 1$.

4. Area under the curve: The overall performance of a classifier over all thresholds is given by the area under the ROC curve (AUC). A good predictive model has AUC closer to 1.

   The AUC of a classifier is the probability that a classifier will rank a randomly chosen instance with $y = 1$ higher than a randomly chosen $y = 0$ instance. A classifier that randomly guesses the class $y = 1$ half the time is expected to get half the labels with $y = 1$ and 0 correctly (the point (0.5,0.5) on the ROC curve). A classifier that randomly guesses the class $y = 1$, 90% of the time is expected to get 90% of the positives correct but FPR will also increase to 90% (the point (0.9,0.9) on the ROC curve). The AUC of a random classifier is hence 0.5.