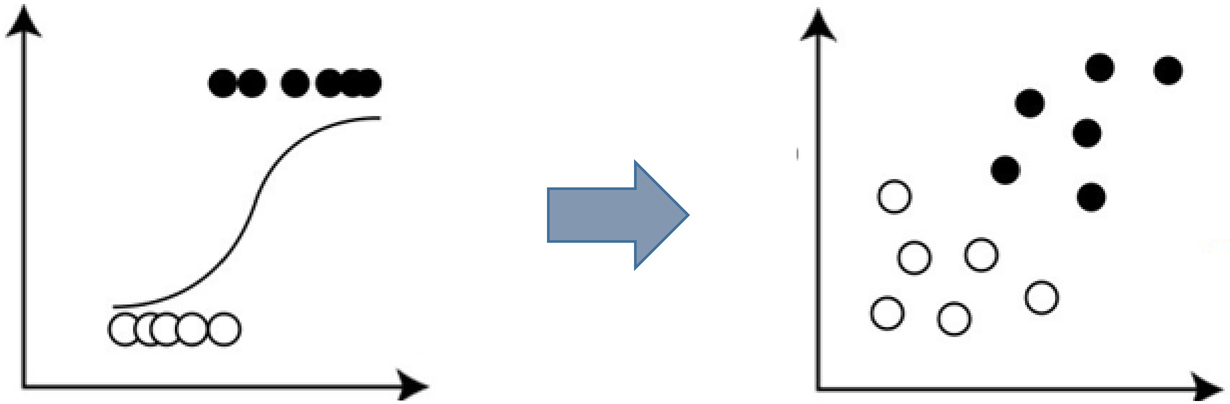
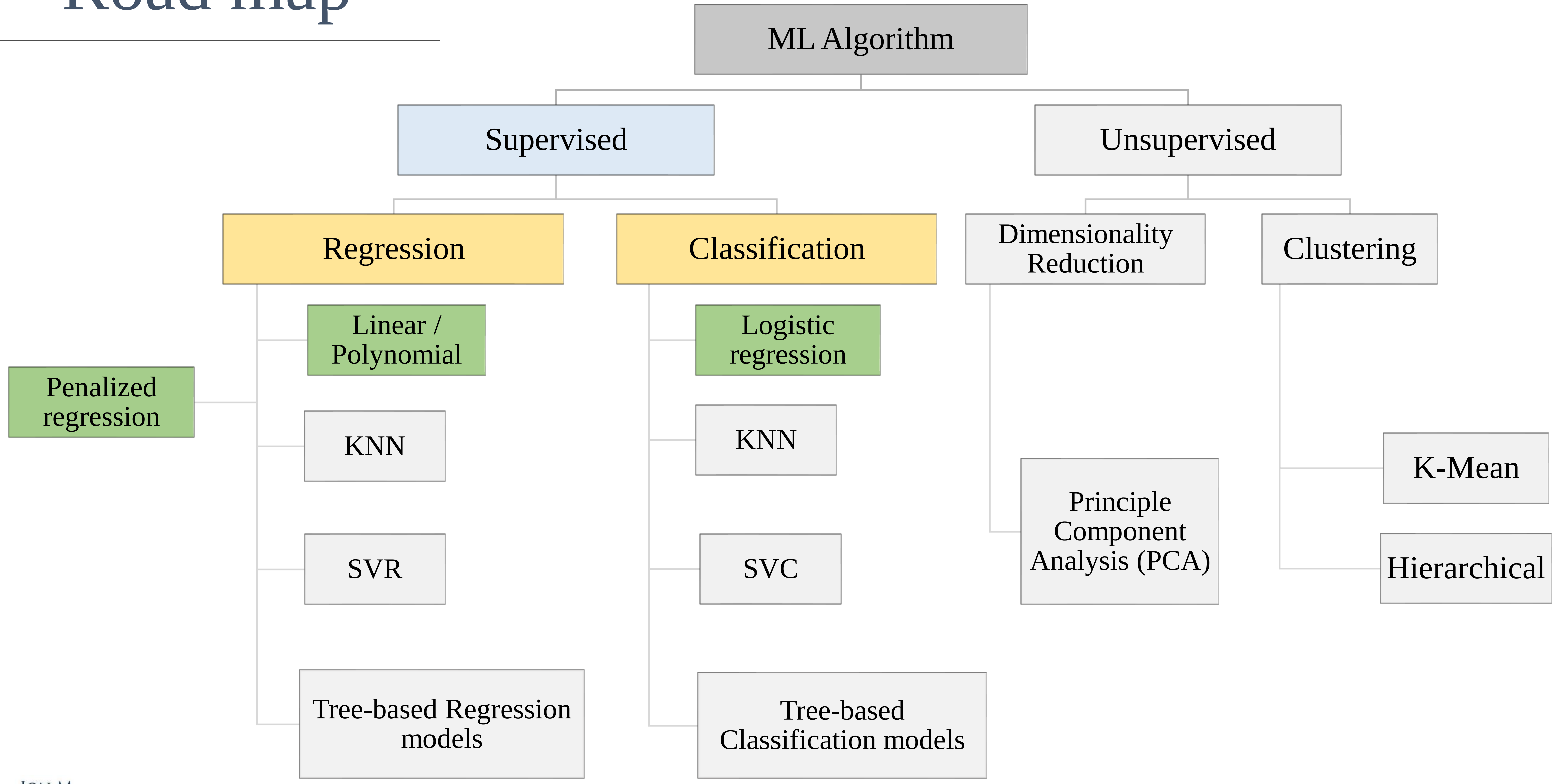
Class 11 – Logistic Regression

Prof. Pedram Jahangiry



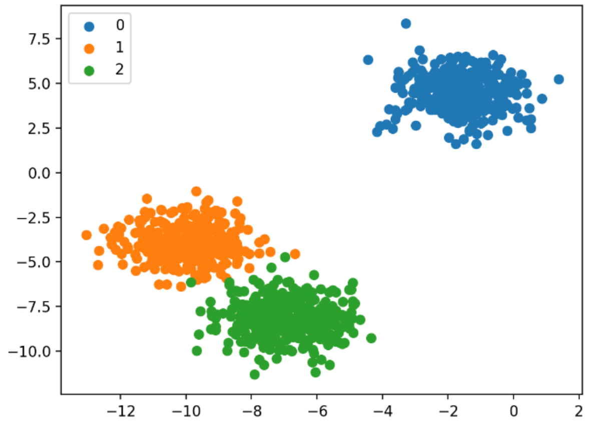
Road map


ML Algorithm
Supervised
Unsupervised
Regression
Classification
Dimensionality Reduction
Clustering
Linear / Polynomial
Penalized regression
KNN
SVR
Tree-based Regression models
Logistic regression
KNN
SVC
Tree-based Classification models
Principle Component Analysis (PCA)
K-Mean
Hierarchical

Topics

1. Linear probability model (LPM) vs Logistic regression
2. Sigmoid function
3. Logistic regression
4. Regularized logistic regression
5. Maximum Likelihood Estimation
6. MLE and GD

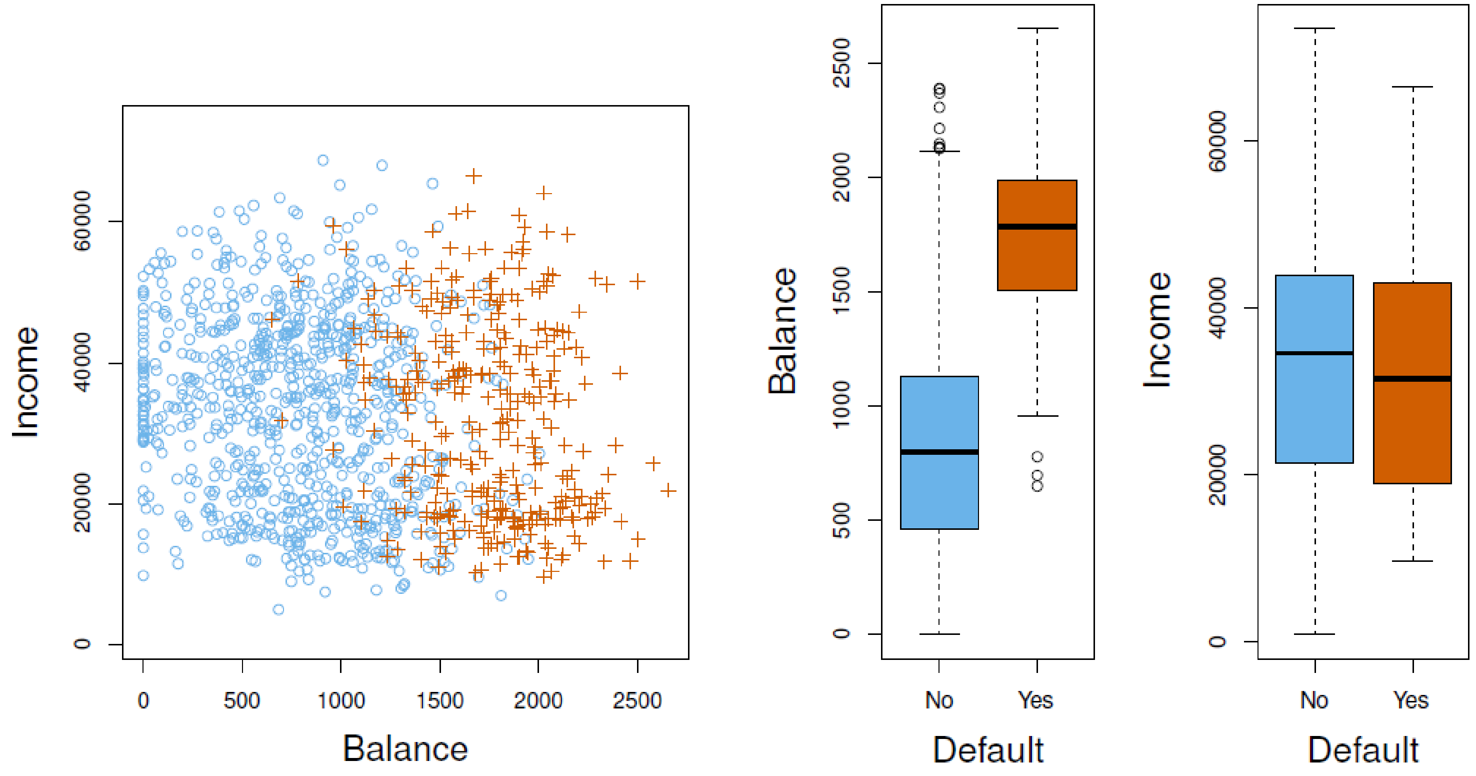
Classification



- Qualitative variables can be either nominal or ordinal.
- Qualitative variables are often referred to as **categorical.**
- **Classification** is the process of predicting categorical variables.
- Classification problems are quite common, perhaps even more than regression problems.
- **Examples:**
  - Financial instrument tranches (investment grade or junk)
  - Online transactions (fraudulent or not)
  - Loan application (approved or denied)
  - Credit card default (default or not)
  - Car insurance customers (high, medium, low risk)

Credit card default example

- Goal: Build a classifier that performs well in both train and test set.

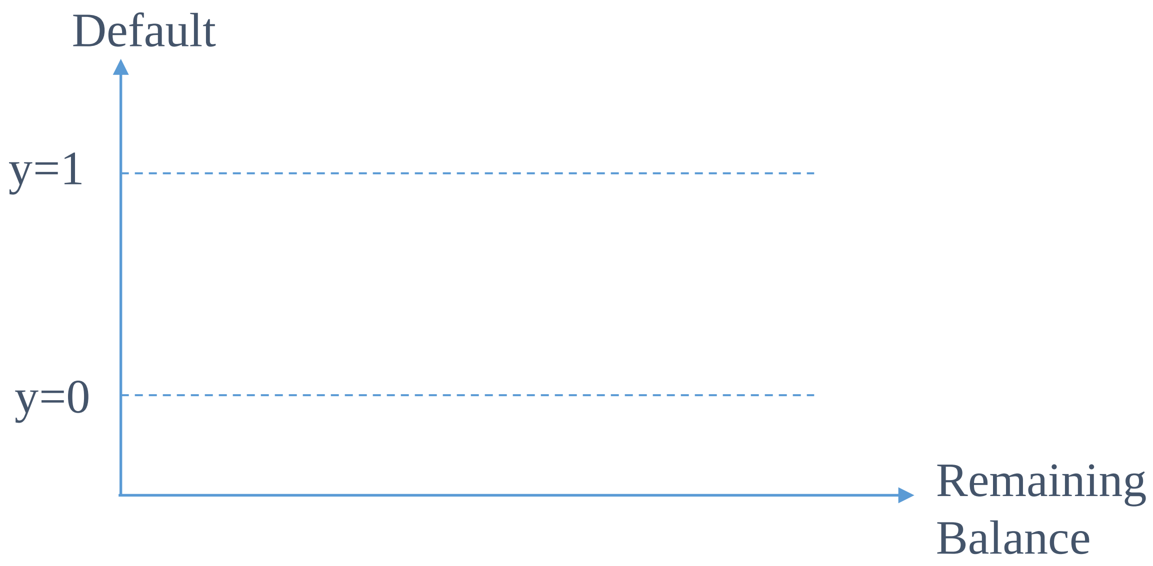


Linear Probability Model (LPM) vs Logistic Regression

Starting with simple LPM : $y = \beta_0 + \beta_1 bal + \epsilon$ where, $Y = 1$ for default and 0 otherwise.

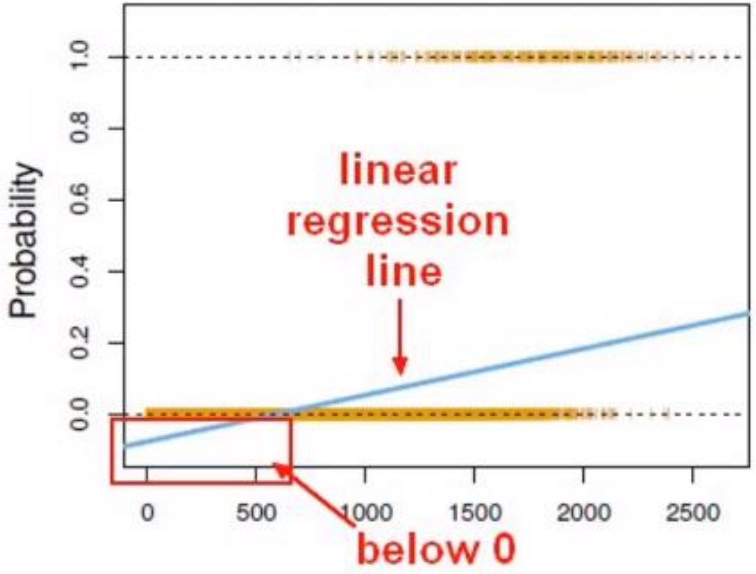
$$E(Y|bal) = \sum P(y_i|bal).\,y_i = \Pr(Y = 1|bal) = P(x) = \beta_0 + \beta_1 bal$$

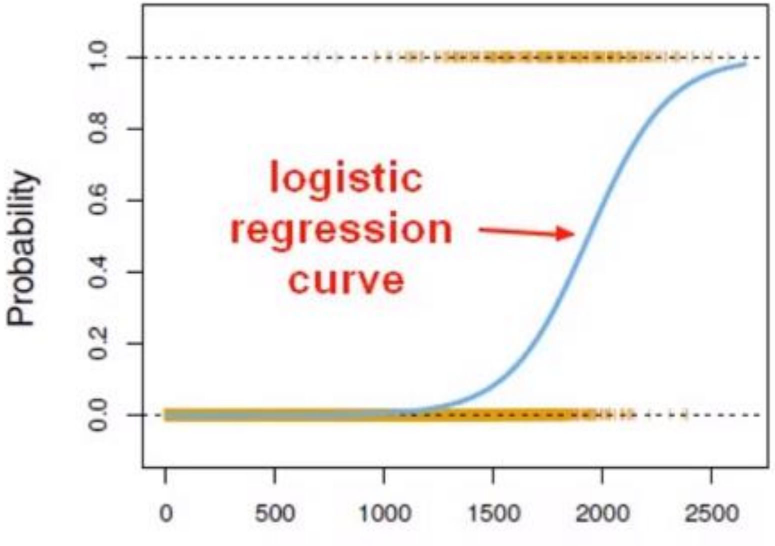
- It seems that simple regression is perfect for this task,
- But what are the caveats?



Linear Probability Model (LPM) vs Logistic Regression

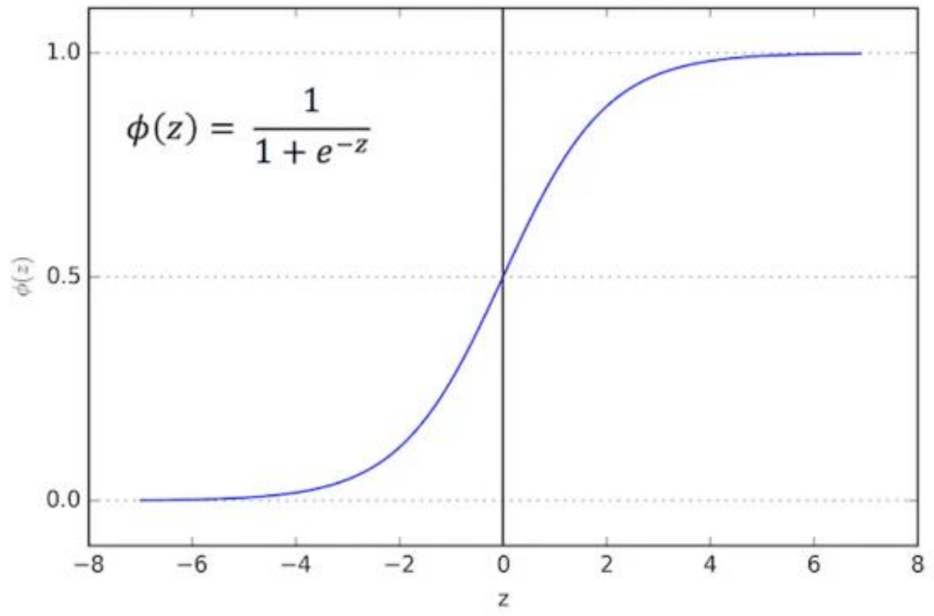
- What else? What if the data set is imbalanced?

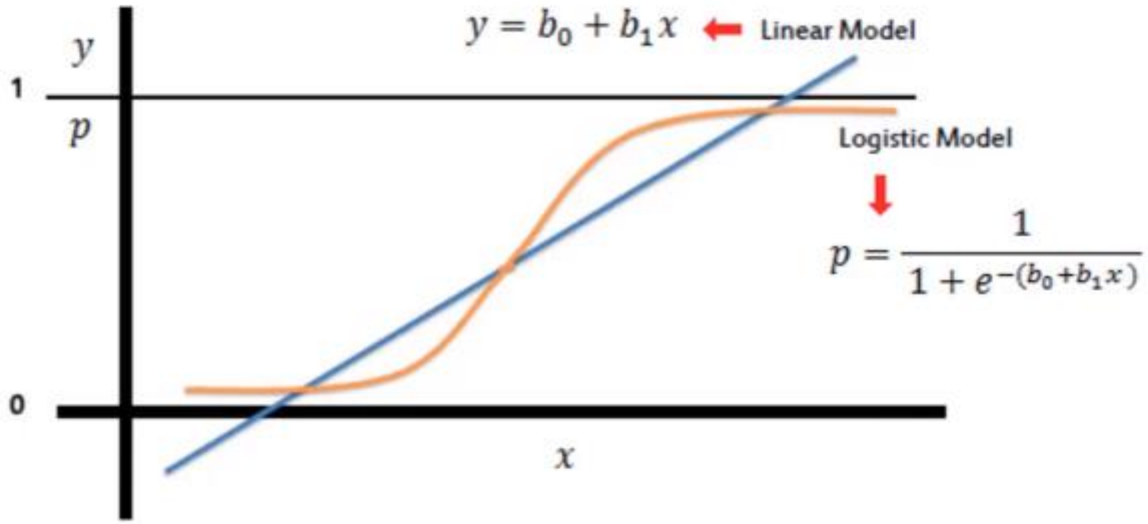




Sigmoid Function

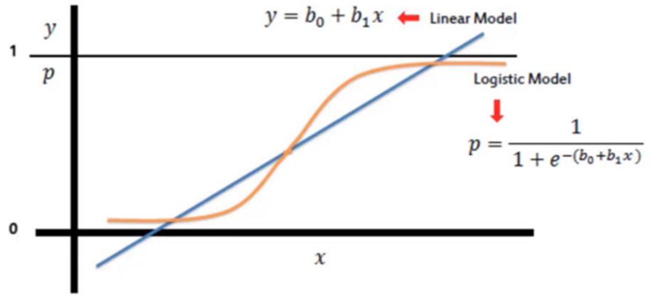
- We need a monotone mapping function that has a range of [0,1]





Logistic Regression (Model)

- The model: $$f_{w,b}(X) = \frac{1}{1+e^{-(WX+b)}}$$



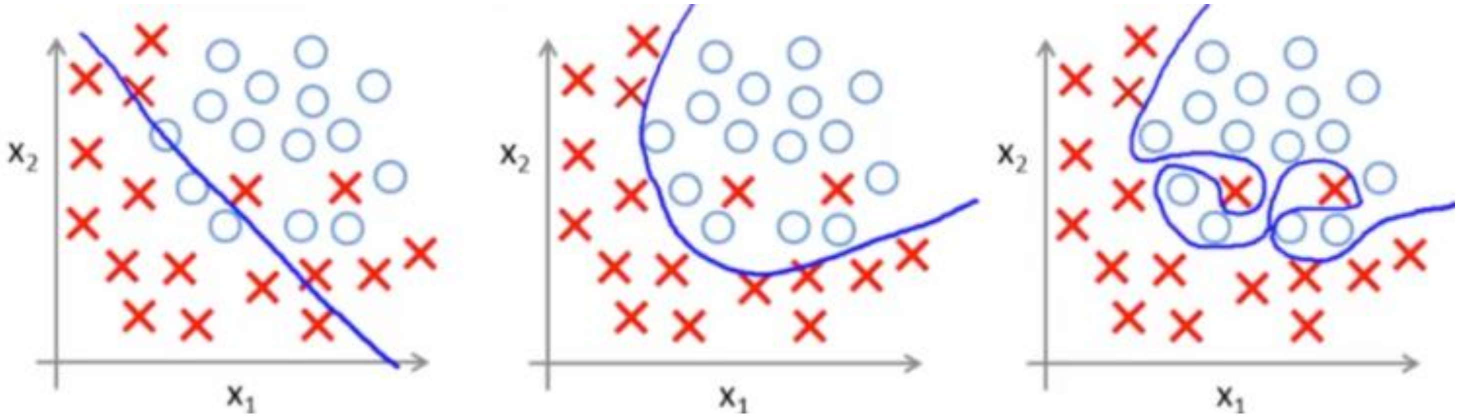
- In case of two classes, $f_{w,b}(X) = \Pr(Y = 1|x) = p(x)$.
- A bit of rearrangement gives

$$Log\left(\frac{p(X)}{1-p(x)}\right) = WX + b$$

- This monotone transformation is called the log odds or logit transformation of $p(x)$.
- Logistic regression ensures that our estimates always lie between 0 and 1

Logistic regression fit (Decision boundary)

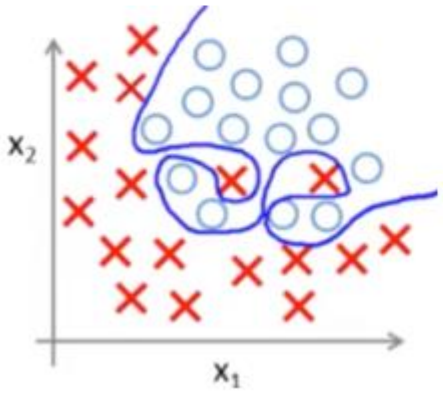
- Depending on how we define $WX + b$, we can get any of the following fits from logistic regression classifier.



Regularized Logistic regression

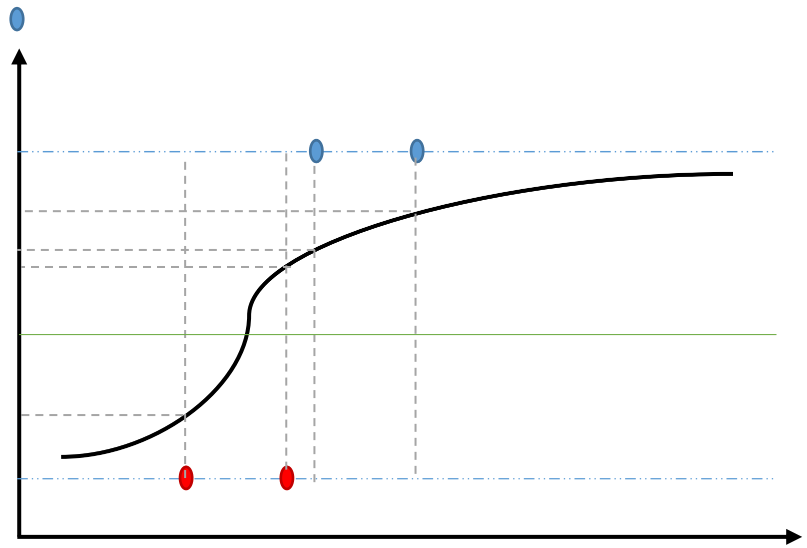
- Penalizing the logistic regression by adding L1, L2 or the combination penalty term.

$$J_{w,b} = \text{logistic cost function} + \lambda_1 \sum_{j=1}^{D} |w_j| + \lambda_2 \sum_{j=1}^{D} w_j^2$$

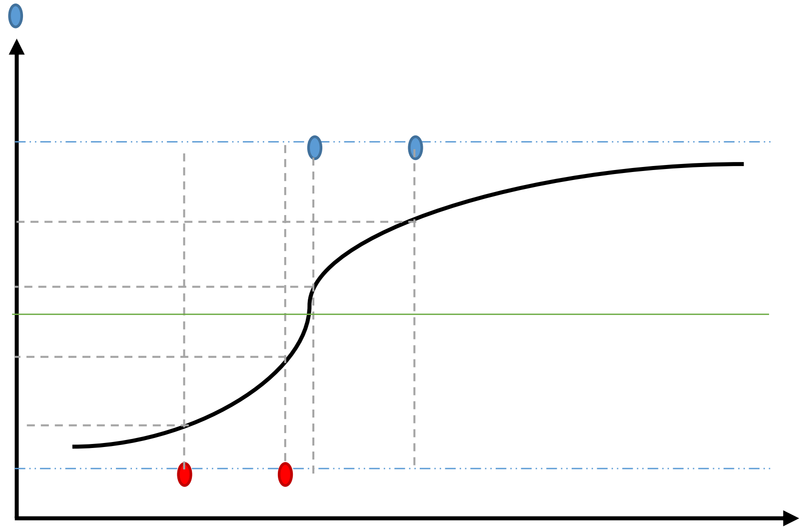


Logistic Regression Estimation (Maximum Likelihood)

- In logistic regression, instead of minimizing the average loss, we maximize the **likelihood** of the training data according to our model. This is called maximum likelihood estimation.
- What is the likelihood function?
- The likelihood function describes the joint probability of the observed data as a function of the parameters of the model.



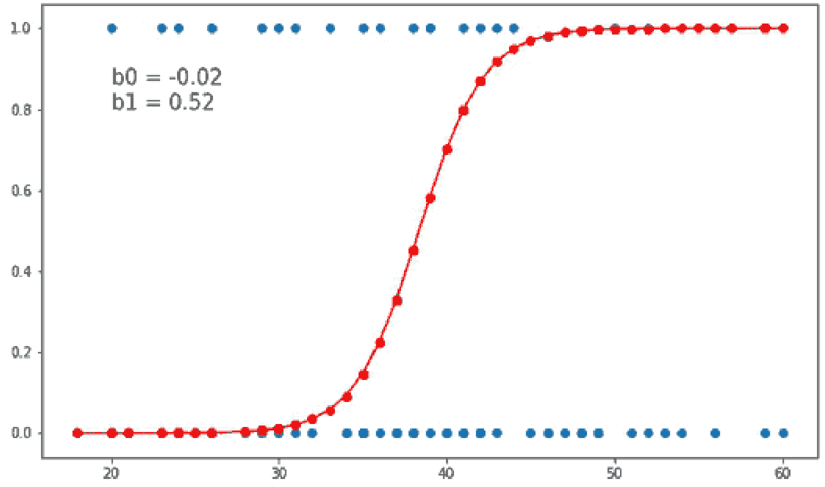
$$L = 0.9 * 0.8 \; * (1 - 0.75) * (1 - 0.2) = 0.144$$



$$L = 0.85 * 0.6 \; * (1 - 0.4) * (1 - 0.2) = 0.244$$

Logistic Regression (Maximum Likelihood)

- MLE in action!



$$f_{w^*,b^*}(X) = \frac{1}{1+e^{-(W^*X+b^*)}}$$

$$L_{w,b} = \prod_{i} f_{w,b}(x_i)^{y_i}\left(1 - \mathrm{f}_{w,b}(x_i)\right)^{1-y_i}$$

Logistic Regression (Objective function)

- Maximizing the likelihood function:

$$Max\ \{L_{w,b} = \prod_i \ f_{w,b}(x_i)^{y_i}\left(1 - f_{w,b}(x_i)\right)^{1-y_i}\}$$

- **Solution**: In practice, it is more convenient to maximize the log-likelihood function. This log-likelihood maximization, gives us w^* and b^*. There is no closed form solution to this optimization problem. We need to use gradient descent.

- We are now ready to make **predictions**.

$$f_{w^*,b^*}(X) = \frac{1}{1+e^{-(W^*X+b^*)}}$$

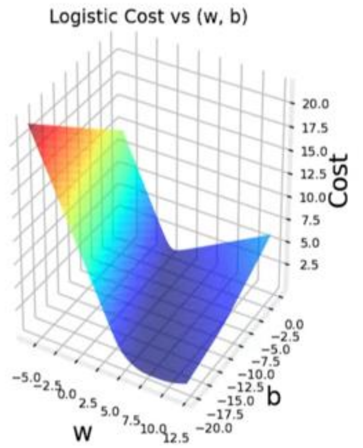
- Depending on how we define the probability threshold, we can classify the observations. In practice, the choice of the threshold could be different depending on the problem.

MLE and Gradient Descent

Simplified loss function

$$L\left(f_{\vec{w},b}\left(\vec{x}^{(i)}\right), y^{(i)}\right) = \begin{cases} -\log\left(f_{\vec{w},b}\left(\vec{x}^{(i)}\right)\right) & \text{if } y^{(i)} = 1 \\ -\log\left(1 - f_{\vec{w},b}\left(\vec{x}^{(i)}\right)\right) & \text{if } y^{(i)} = 0 \end{cases}$$

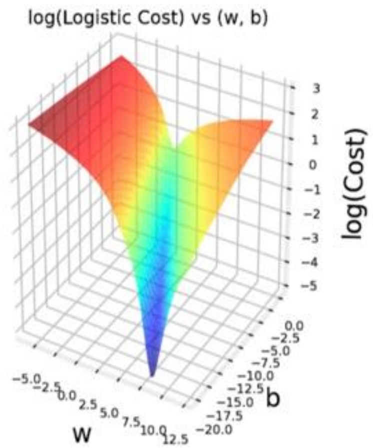
$$\boxed{L\left(f_{\vec{w},b}\left(\vec{x}^{(i)}\right), y^{(i)}\right) = -y^{(i)}\log\left(f_{\vec{w},b}\left(\vec{x}^{(i)}\right)\right) - \left(1 - y^{(i)}\right)\log\left(1 - f_{\vec{w},b}\left(\vec{x}^{(i)}\right)\right)}$$



cost

$$J(\vec{w}, b) = \frac{1}{m}\sum_{i=1}^{m}\left[\underbrace{L\left(f_{\vec{w},b}\left(\vec{x}^{(i)}\right), y^{(i)}\right)}\right]$$

$$= -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log\left(f_{\vec{w},b}\left(\vec{x}^{(i)}\right)\right) + \left(1 - y^{(i)}\right)\log\left(1 - f_{\vec{w},b}\left(\vec{x}^{(i)}\right)\right)\right]$$



Logistic regression output for credit card default example

$$P(default|bal, inc) = \frac{1}{1 + e^{-(b + w_1(bal) + w_2(inc))}}$$

		Predictions (Decision boundary)	
		0 No Default	1 Default
Actual	0 No Default	TN=1933	FP=3
	1 Default	FN=44	TP=20

Jon M. Huntsman School of Business
Utah State University


class
DISMISSED