1. Classification

$\mathbb{R}^n \ni X \rightsquigarrow Y \in \mathbb{R} \leftarrow$ quantitative

$\mathbb{R}^n \ni X \rightsquigarrow Y \in \{0, 1\}$ qualitative

- Classyfying is a process of assigning observation to categories
- In the classification problem we work with qualitative responses instead of quantitative

sklearn, datasets. load_boston

sklearn, datasets, load_iris

- Often we want to predict the probability that observation belongs to each of the categories
- Among classification techniques (classifier) we have
  - logistic regression
  - linear discriminant analysis (LDA)
  - quadratic —"— (QDA)
  - K- nearest neighbors
  - tree, random forests
  - SVM (Support Vector Machines)

2. Why not linear regression?

- there is a problem with converting a qualitative response variable with more than 2 classes into a qualitative response that can be used in linear regession
- there is a problem with obtaining reasonable estimates of $\mathbb{P}(Y=1 \mid X)$

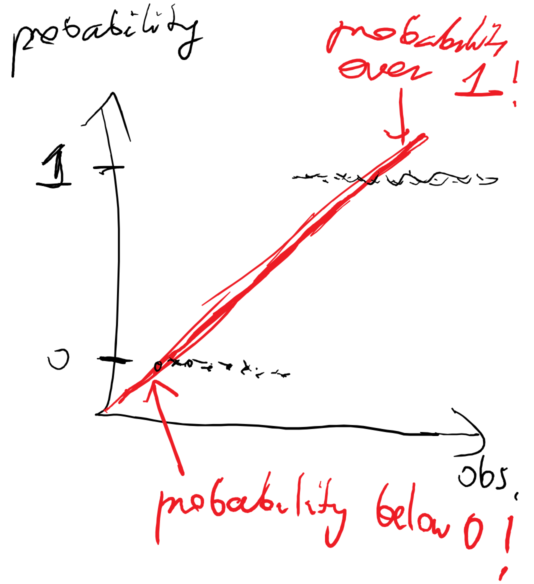
$$Y = \begin{cases} 0 & \text{if cat} \\ 1 & \text{if dog} \\ 2 & \text{if bird} \end{cases}$$

$$Y^{(1)} = \begin{cases} 0 & \text{if not cat} \\ 1 & \text{if cat} \end{cases}$$

$$Y^{(2)} = \begin{cases} 0 & \text{if not dog} \\ 1 & \text{if dog} \end{cases}$$

$$\cancel{Y^{(3)} = \begin{cases} 0 & \text{if not bird} \\ 1 & \text{if bird} \end{cases}}$$

not necessary



3. Logistic Regression – Model

binary classification $\quad \mathbb{R}^n \ni X \rightsquigarrow Y \in \underbrace{\{0,1\}}^{\text{classes/categories}}$

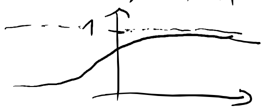
(predictors) $\quad$ (responses)

Model

$$p(x) = \mathbb{P}(Y=1 \mid X=x)$$
$$= \varphi(\beta_0 + \beta_1 x)$$
$$= \frac{1}{1+e^{\beta_0+\beta_1 x}}$$
$$= \frac{e^{\beta_0+\beta_1 x}}{e^{\beta_0+\beta_1 x}+1}$$

$\varphi : \mathbb{R} \to [0,1] \quad \varphi \nearrow$

$\varphi(-\infty)=0, \; \varphi(\infty)=1$



$$\varphi(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1}$$

(logistic function/sigmoid)

Hence

$$e^{\beta_0+\beta_1 x} = p(x)\left(1+e^{\beta_0+\beta_1 x}\right)$$
$$e^{\beta_0+\beta_1 x}(1-p(x)) = p(x)$$
$$\frac{p(x)}{1-p(x)} = e^{\beta_0+\beta_1 x}$$
$$\underbrace{\log\left(\underbrace{\frac{p(x)}{1-p(x)}}_{\text{odds}}\right)}_{\text{log odds (logit)}} = \beta_0+\beta_1 x$$

Thus logistic regression has a logit (log odds) that is linear in X,

4. Logistic Regression - fitting

- to fit the model given by $p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$ we use <u>maximum likelihood</u> method:

$$\ell(\beta_0, \beta_1) := \prod_{i:\, y_i=1} p(x_i) \prod_{i':\, y_{i'}=0} (1-p(x_{i'}))$$

$p(x) = P(Y=1|X)$

- the estimates of β_0, β_1 are chosen to maximize this likelihood function $\ell(\beta_0, \beta_1)$
- intuiton: we try to $\hat{\beta}_0, \hat{\beta}_1$ such that if we put them into $p(x)$ we get number close to 1 for each observations such that $Y=1$ and we get number close to 0 for all observations that $Y=0$.

5. Logistic Regression – multiple case

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

  where $X = (X_1, \dots, X_p)$ – predictors.

- Hence
  $$p(X) = \frac{e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}}$$

- We use maximum likelihood method to predict $\beta_0, \beta_1, \dots, \beta_p$.

6. Logistic Regression – multinomial case

- we can extend two-classes (binary classification) logistic regression to the case of $K > 2$ classes
- At first, we select a single class to be a baseline, e.g. let it be Kth class.
- Then, we replace the model $p(x) = \frac{e^{\beta_0 + \beta_1 x + \dots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}}$ with the model

$$\mathbb{P}(Y = j \mid X = x) = \frac{e^{\beta_{j0} + \beta_{j1} x_1 + \dots + \beta_{jp} x_p}}{1 + \sum_{l=1}^{K-1} e^{\beta_{l0} + \beta_{l1} x_1 + \dots + \beta_{lp} x_p}}$$

$$\mathbb{P}(Y = K \mid X = x) = \frac{1}{1 + \sum_{l=1}^{K-1} e^{\beta_{l0} + \beta_{l1} x_1 + \dots + \beta_{lp} x_p}}$$

- Then

$$\log\left(\frac{\mathbb{P}(Y = j \mid X = x)}{\mathbb{P}(Y = K \mid X = x)}\right) = \beta_{j0} + \beta_{j1} x_1 + \dots + \beta_{jp} x_p$$

- choosing Kth class as the baseline is not important in the sense that we can choose any other class as a baseline

7. Logistic Regression – softmax

- soft-max is an alternative coding for multinomial logistic regression.
- Instead selecting baseline class, we treat all K classes symmetrically and we assume

$$\mathbb{P}(Y=j \mid X=x) = \frac{e^{\beta_{j0}+\beta_{j1}x_1+\ldots+\beta_{jp}x_p}}{\sum_{l=1}^{K} e^{\beta_{l0}+\beta_{l1}x_1+\ldots+\beta_{lp}x_p}}$$

- Thus, we estimate coefficients for all K classes and we get

$$\log\left(\frac{\mathbb{P}(Y=j \mid X=x)}{\mathbb{P}(Y=j' \mid X=x)}\right) = (\beta_{j0}-\beta_{j'0}) + (\beta_{j1}-\beta_{j'1})x_1 + \ldots + (\beta_{jp}-\beta_{j'p})x_p$$

8. Logistic Regression – remarks

- In a linear regression ($Y = \beta_0 + \beta_1 X$) model β_1 gives the average change in Y asociated with increasing of X by one unit
- In a Logistic Regression model ($\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 X$, where $p(x) = \mathbb{P}(Y=1 \mid X=x)$) increasing X by one unit changes the log odds ($\log\left(\frac{p(x)}{1-p(x)}\right)$) by β_1.
- Logistic regression (for binary classification; $p(x) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$) corresponds to neural net with one neuron and activation function $\varphi(x) = \frac{1}{1+e^{-x}}$ and cost function $-l(\beta_0, \beta_1)$