University of Essex

# Classification Problem

Akitaka Matsuo
Essex IADS

# Content

University of Essex

- Define classification problem
- Performance criteria
- Classification algorithms:
  - Logistic regression
  - k-Nearest Neighbors (KNN)
  - Others

# Two types of machine learning problem

University of Essex

- Regression problem:
  - The outputs are continuous
    - We saw KNN regression / OLS / regularized regression
- Classification problem:
  - The outputs are categorical

# Classification problem

University of Essex

- $Y$ is a categorical variable
  - Example:
    - Vote for a Republican candidate or not (binary)
    - Choice of college major (multinomial)
- Inputs $X$ can be anything
  - continuous
  - categorical
- This lecture discusses a binary classification problem where only two outcome categories
  - feasibility of advanced model
  - multinomial = combining multiple binary choices

# Performance criteria

University of Essex

- Accuracy
- f1
  - Precision
  - Recall
- Area under curve of ROC

# Confusion matrix

University of Essex

– Cross-tabulation of true $Y$ and predicted $Y$
  - TP: True Positive
  - TN: True Negative
  - FP: False Positive
  - FN: False Negative

| | | Predicted class: Positive | Predicted class: Negative |
|---|---|---|---|
| Actual class | Positive | TP | FN |
| | Negative | FP | TN |

# Accuracy

University of Essex

- How the model correctly predicts $Y$
- $\frac{\mathrm{TP}+\mathrm{TN}}{N}$
  - $N$ is number of obs
- Not reliable measure if imbalanced classes
  - Example:
    - Actual positive: 90
    - Actual Negative: 10

| | | Predicted | |
|---|---|---|---|
| | | positive | negative |
| Actual | positive | TP: 35 | FN: 15 |
| | negative | FP:10 | TN: 40 |

# Precision

University of Essex

- Class specific measure (i.e. we need to determine which class is positive category)
- If the prediction is Positive. How likely it is true?
- $\frac{\mathrm{TP}}{\mathrm{TP+FP}}$
- In this example: $\frac{35}{35+10} = .778$

| | | Predicted | |
|---|---|---|---|
| | | positive | negative |
| Actual | positive | TP: 35 | FN: 15 |
| | negative | FP:10 | TN: 40 |

# Recall

University of Essex

- Again, class specific measure
- Among the positive cases, how likely
- $\frac{\text{TP}}{\text{TP+FN}}$
- In this example: $\frac{35}{35+15} = .70$

| | | Predicted | |
|---|---|---|---|
| | | positive | negative |
| Actual | positive | TP: 35 | FN: 15 |
| | negative | FP:10 | TN: 40 |

# F1

University of Essex

- A harmonic average of precision and recall
- 2 * (Precision * Recall) / (Precision + Recall)
- Domain: [0, 1]
- In this example: $\frac{.778*.7*2}{.778+.7} = 0.737$

| | | Predicted | |
|---|---|---|---|
| | | positive | negative |
| Actual | positive | TP: 35 | FN: 15 |
| | negative | FP:10 | TN: 40 |

# Precision, recall, F1 in imbalanced

University of Essex

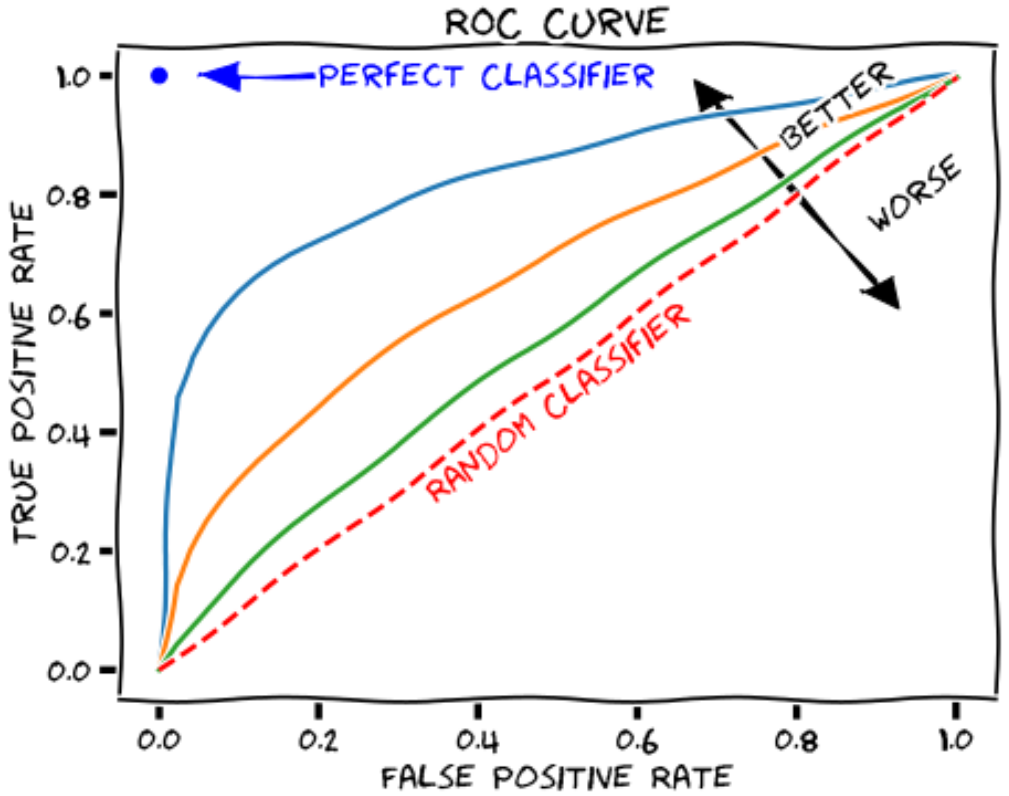
- Accuracy: 0.91
- Positive class
  - Precision: 1
  - Recall: 0.1
  - F1: 0.182

| | | Predicted | |
|---|---|---|---|
| | | positive | negative |
| Actual | positive | TP: 1 | FN: 9 |
| | negative | FP: 0 | TN: 90 |

# ROC, Area under the curve

University of Essex

- ROC: Receiver operating characteristic
- AUC: Area under the curve of ROC
  - min = 0.5
  - max = 1
- How much improvement of TP without sacrificing FP




https://commons.wikimedia.org/wiki/File:Roc-draft-xkcd-style.svg

# Logistic Regression

University of Essex

- Logistic regression is one of the standard methods for classification problem.
- The model determines the probably of positive by converting a linear function to a probability.
- The linear function is:

$$f(X) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p = \beta_0 + \sum_j \beta_j X_j$$

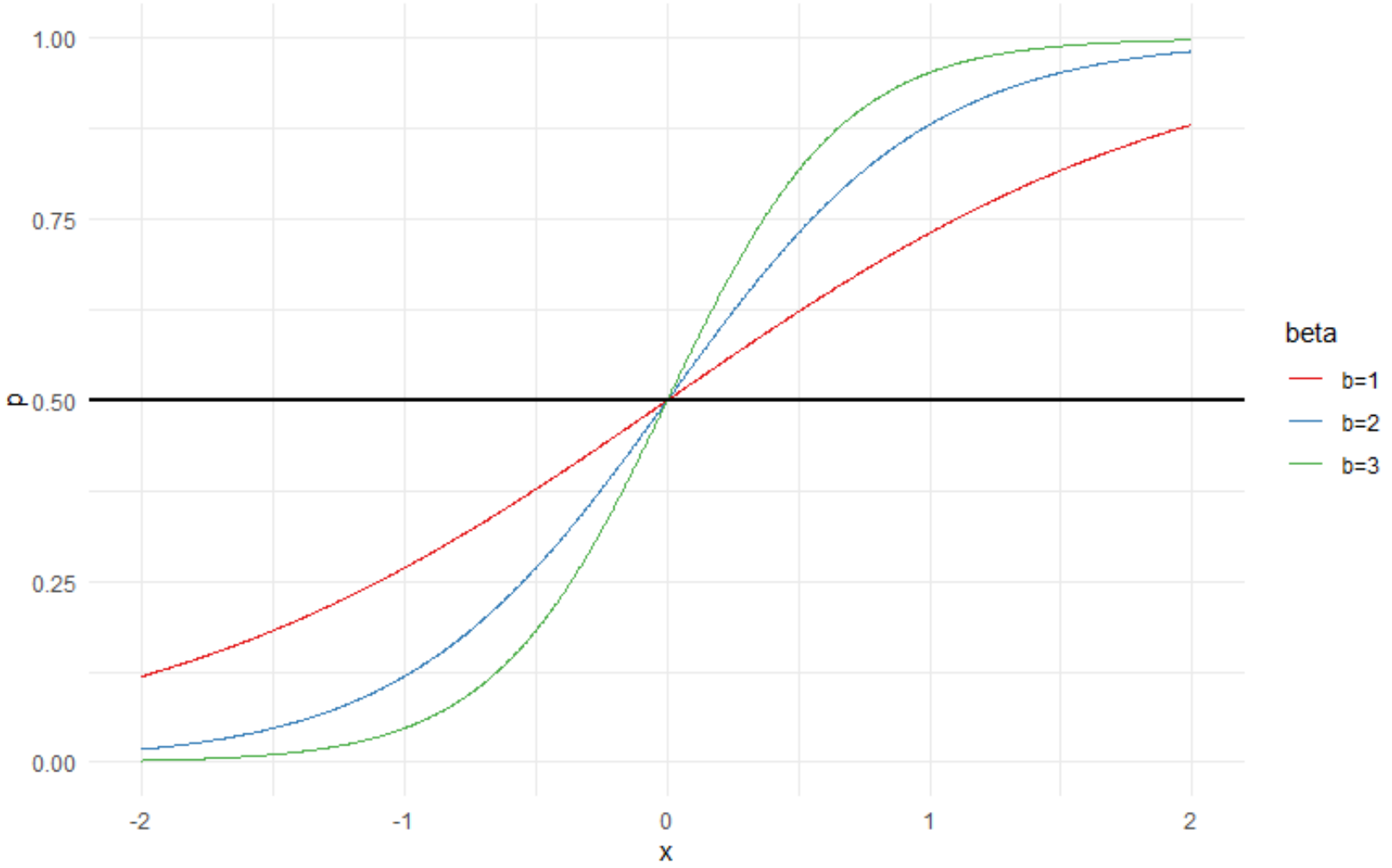
- The conversion is done through logistic function (so it's called ligistic regression)

$$p(X) = \frac{e^{f(X)}}{1 + e^{f(X)}} = \frac{e^{\beta_0 + \sum_j \beta_j X_j}}{1 + e^{\beta_0 + \sum_j \beta_j X_j}}$$

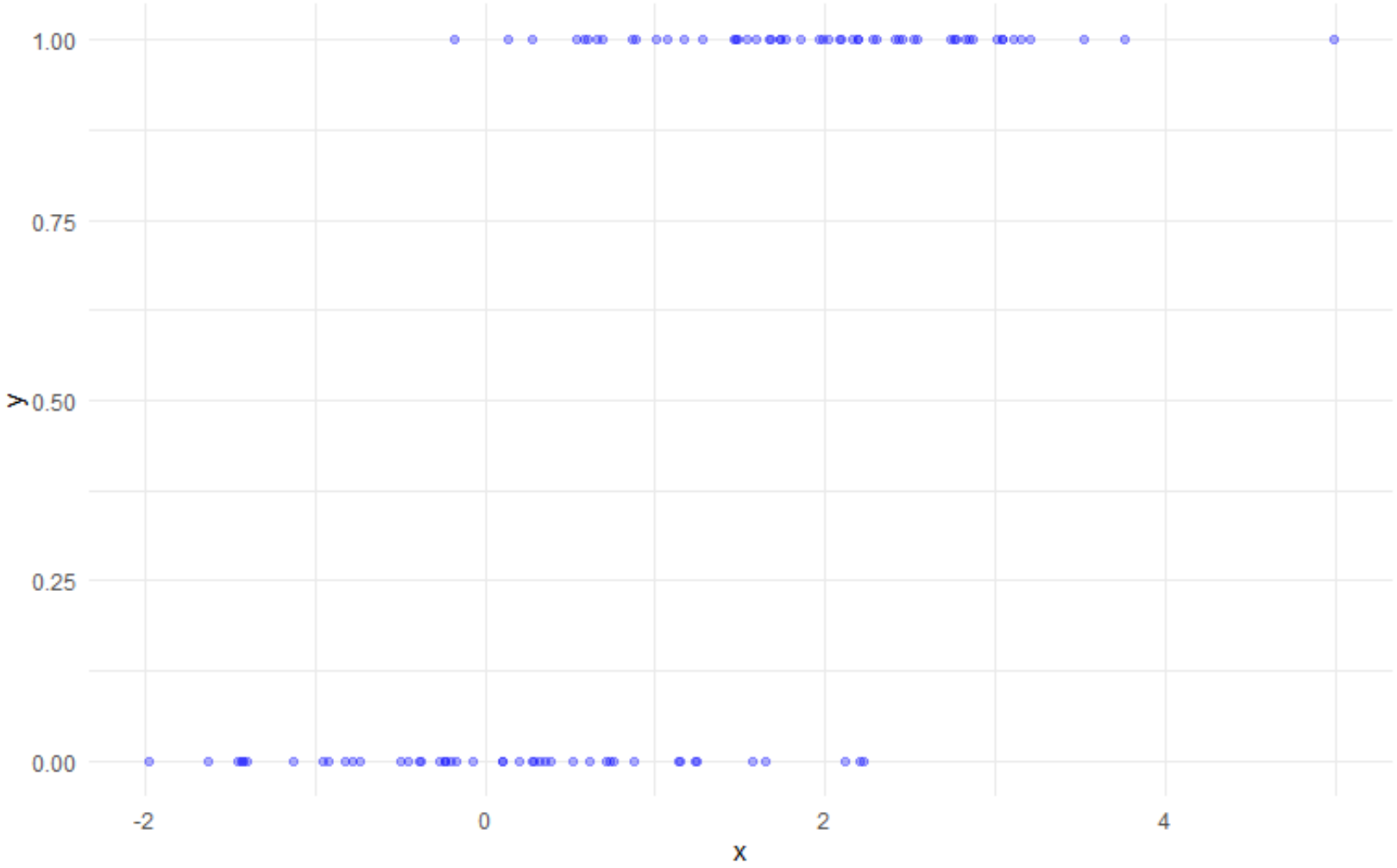
- When $f(x)$ increses $p(X)$ increases
- When $f(x) = -\infty \rightarrow p(X) = 0$
- When $f(x) = \infty \rightarrow p(X) = 1$
- There is no tuning parameter

# Logistic regression illustration

University of Essex

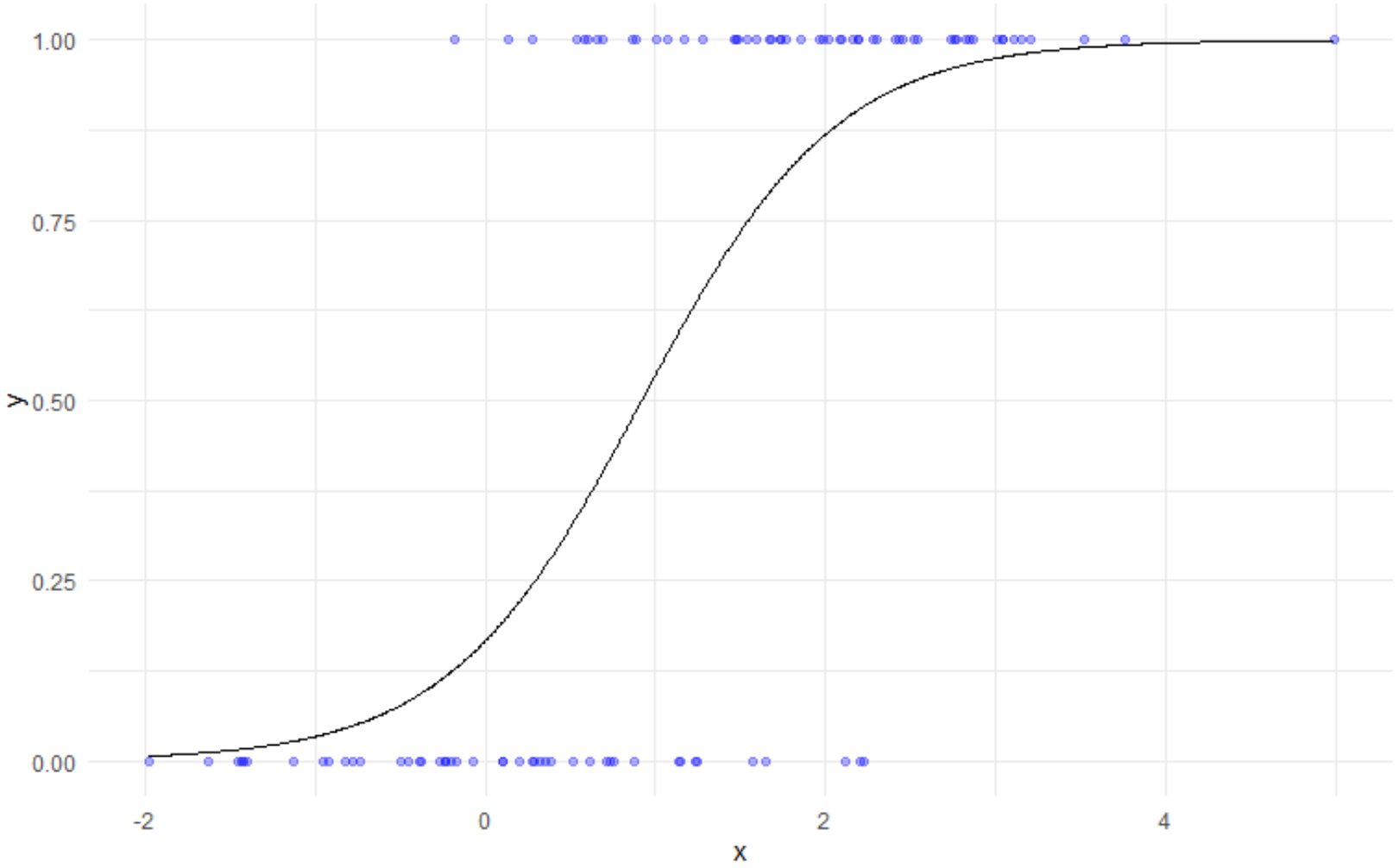


# Logistic regression illustration: data example



# Logistic regression illustration: Plot

University of Essex



# Example of Other Classification Methods

University of Essex

- KNN Classifier
- Tree based methods
  - Random forest
  - Bagging/Boosting
- Support vector machine (SVM)
  - Choice of kernels
- Neural Network (Deep learning)
  - Go beyond sklearn
  - TensorFlow, pytorch