



Linear Classification

Réda DEHAK
<http://ismil.dehak.org>

1






Contents

- Classification
- Linear Classification
- Logistic Regression
- Performance metrics

Réda DEHAK 2

2






Classification

- learning a **function** (model) that **maps** an **input** (features) vectors to a **discret output** (target or labels) based on **examples input-output pairs**.
- Examples:
 - Binary classification: $y \in \{0, 1\}$
 - Email : Spam / Not Spam?
 - Online Transactions: Fraudulent (Yes/No)
 - Check Identity: Target / impostors
 - Multiclass classification:

Réda DEHAK 3

3



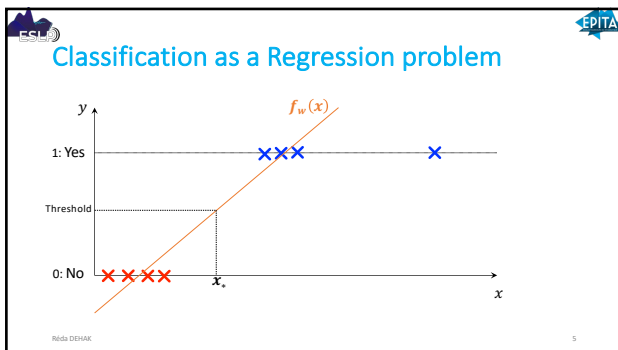


Classification

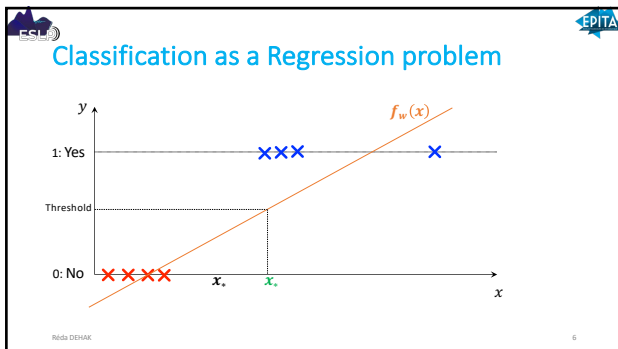
- Logistic Regression
- Support Vector Machines (SVM)
- Neural Networks
- K-Nearest Neighbors (KNN)
- Naïve Bayes
- Decision Trees
-

Réda DENAE
4

4



5



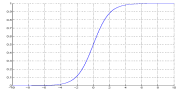
6

Logistic Regression

- Want $0 \leq f_W(x) \leq 1$:
 - Linear Regression: $f_W(x) = W^T x$
 - Logistic Function: $h(z) = \frac{1}{1 + e^{-z}}$

$z = h^{-1}(x) = \text{logit}(x) = \log\left(\frac{x}{1-x}\right)$

$\frac{dh(z)}{dz} = h(z)(1-h(z))$



7

Logistic Regression

- Want $0 \leq f_W(x) \leq 1$:
 - Logistic Regression: $P(y=1) = f_W(x) = \frac{1}{1 + e^{-W^T x}}$
 - $\frac{df_W(x)}{dw} = x f_W(x)(1 - f_W(x))$
 - $P(y=0) = 1 - P(y=1)$
- Decision Function:
 - $P(y=1) \geq P(y=0)$: Positive Class
 - Otherwise: Negative Class

8

Logistic Regression

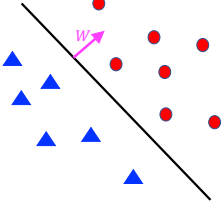
- $P(y=1) \geq P(y=0) \Leftrightarrow \frac{P(y=1)}{P(y=0)} \geq 1$
 - $\Leftrightarrow \log\left(\frac{P(y=1)}{P(y=0)}\right) \geq 0$
 - $\Leftrightarrow \log\left(\frac{P(y=1)}{1-P(y=1)}\right) = \text{logit}(P(y=1)) \geq 0$
 - $\Leftrightarrow \text{logit}(P(y=1)) = \mathbf{W}^T \mathbf{x}$

9

HyperPlane

- Definition:
 $W^T x + b = 0$
- Divide the space into two parts:
 - 1) $W^T x + b > 0$: direction of the vector w (red in the figure)
 - 2) $W^T x + b < 0$: opposite of the direction of w (blue in the figure)
- Distance to the hyperplane:

$$\frac{|W^T x + b|}{\|W\|}$$



10

Logistic Regression

- $P(y = 1) \geq P(y = 0) \Leftrightarrow \frac{P(y=1)}{P(y=0)} \geq 1$
 $\Leftrightarrow \log \left(\frac{P(y=1)}{P(y=0)} \right) \geq 0$
 $\Leftrightarrow \log \left(\frac{P(y=1)}{1-P(y=1)} \right) = \text{logit}(P(y = 1)) \geq 0$
 $\Leftrightarrow \text{logit}(P(y = 1)) = \mathbf{w}^T \mathbf{x} \geq 0$

Boundary = Hyperplane

11

Linear Classification

- Classification : Decision Function is a hyperplane in input space.
- Classification methods:
 - Logistic Regression
 - Perceptron
 - SVM
 - Naive Bayes

12

Training: Two Classes

$$\mathcal{X} = \{(x_i, y_i), i = 1 \dots N\} \quad y_i \in \{0, 1\}$$

$y_i \sim \text{Bernoulli}$

- Score Function : Maximum LLK (Log Likelihood)
- $LLK(\mathcal{X}) = \prod_{i=1}^N P(y_i | x_i)$
- $= \prod_{i=1}^N (f_W(x_i))^{y_i} (1 - f_W(x_i))^{1-y_i}$
- $LLK(\mathcal{X}) = \sum_{i=1}^N y_i \log(f_W(x_i)) + (1 - y_i) \log(1 - f_W(x_i))$

Réda DENAE 13

13

Training: Two Classes

- $LLK(\mathcal{X}) = \sum_{i=1}^N y_i \log(f_W(x_i)) + (1 - y_i) \log(1 - f_W(x_i))$
- $\frac{dLLK(\mathcal{X})}{dW} = \sum_{i=1}^N x_i y_i \frac{f_W(x_i)(1-f_W(x_i))}{f_W(x_i)} - x_i(1 - y_i) \frac{f_W(x_i)(1-f_W(x_i))}{1-f_W(x_i)}$

$$\frac{dLLK(\mathcal{X})}{dW} = \sum_{i=1}^N x_i (y_i - f_W(x_i))$$

Réda DENAE 14

14

Gradient Descent for linear Regression


$$E = \frac{1}{N} (Y^T Y - 2 A^T X Y + A^T X X^T A)$$

$$\nabla E = \frac{2}{N} (X X^T A - X Y)$$

$$\nabla E = \frac{2}{N} X (X^T A - Y)$$

Réda DENAE 15

15




Multiclass

- Use softMax rather than logistic function:

$$P(y = i|x) = f_{W_{1..k}}(x; i) = \frac{e^{W_i^T \hat{x}}}{\sum_{l=1}^k e^{W_l^T \hat{x}}}$$

Réda DENAE 16

16



No Linear Logistic Regression

- Using a non Linear Mapping φ before the regression.
- Examples: *Quadratic mapping*

$$\varphi \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1^2 \\ x_2^2 \\ x_1 x_2 \\ x_1 \\ x_2 \\ 1 \end{pmatrix}$$

Réda DENAE 17

17



Performance Metrics

Réda DENAE 18

18

Performances: Matrix confusion

- A binary classifier classifies data points as + or -
- If we also know the true classification, the performance of the classifier is a 2×2 contingency table, called a Confusion Matrix.

		Actual Class	
		+	-
Predicted Class	+	True Positives (TP)	False Positives (FP)
	-	False Negatives (FN)	True Negatives (TN)

Réda DENAE 19

19

Performances: Missclassification Errors

- Two types of Errors:
 - False Positives (FP): Type-I errors
 - False Negatives (FN): Type-II errors

		Actual Class	
		+	-
Predicted Class	+	True Positives (TP) Good!	False Positives (FP) Bad! (Type-I errors)
	-	False Negatives (FN) Bad! (Type-II errors)	True Negatives (TN) Good!

Réda DENAE 20

20

Example:

Logistic Regression

		Actual Class	
		+	-
Predicted Class	+	72	29
	-	33	31

SVM

		Actual Class	
		+	-
Predicted Class	+	94	37
	-	11	23

Which is the Best Classifier?

Réda DENAE 21

21

Threshold

- $P(y = 1) \geq P(y = 0)$:

$$P(y = 1) = f_W(x) = \frac{1}{1 + e^{-W^T x}} \geq \mathbf{0.5}$$
- Threshold depends on prior and cost (risk) function

EPITA

22

Threshold

Using costs values to **adjust Threshold** to **minimize costs**

		Actual Class		Score ≥ 0.6		Score ≥ 0.7	
		+	-	+	-	+	-
Predicted Class	+	True Positives (TP) Cost = 0	False Positives (FP) Cost = 20	40	20	30	10
	-	False Negatives (FN) Cost = 5	True Negatives (TN) Cost = 0	10	30	20	40

EPITA



23

Adjust Threshold

Most classifiers have a “knob” or threshold that you can adjust: How certain do they have to be before they classify a “+”? To get more TP’s, you have to let in some FP’s!

EPITA

24

Threshold?

- Notice there is just one free parameter, think of it as TP, since:
 - $FP(TP) = [given\ by\ algorithm]$
 - $TP + FN = P$ (fixed number of actual positives, column marginal)
 - $FP + TN = N$ (fixed number of actual negatives, column marginal)



So all scalar measures of performance are functions of one free parameter (i.e., curves)

The points on any such curve are in 1-to-1 correspondence with those on any other such curve.

- If you ranked some classifiers by how good they are, you might get a different rankings at different points on the scale.
- On the other hand, one classifier might dominate another at all points on the scale.

Réda DEHAË
25

25






Performance Metrics

- Accuracy
- True Positive Rate, Recall, Sensitivity
- False Alarm Rate, False Positive Rate
- Missed Detection Rate, False Negative Rate
- Specificity, True Negative Rate
- Negative Predictive Value
- Precision, Positive Prediction Value
- False Discovery Rate
- F-Score
- F-Measure

Réda DEHAË
26

26

Performance Metrics

- Different combinations of ratios have been given various names. All vary between 0 and 1.
- A performance curve picks one as the independent variable and looks at another as the dependent variable.

actual	
TP	FP
FN	TN
classified	

accuracy (ACC)

actual	
TP	FP
FN	TN
classified	

neg. predictive value (NPV)

actual	
TP	FP
FN	TN
classified	

specificity (SPC)

actual	
TP	FP
FN	TN
classified	

ROC curve

actual	
TP	FP
FN	TN
classified	

precision-recall curve

one minus

one minus

one minus

one minus

Réda DEHAË
27

27

Performance Metrics: Accuracy

Accuracy is the most intuitive performance measure and it is simply a ratio of **correctly predicted** observation to the **total observations**

- $Accuracy = \frac{\text{Total of good Classification}}{\text{Number of Examples}}$
- $Accuracy = \frac{(TP+TN)}{(P+N)}$
- *Accuracy can be a misleading metric for imbalanced data sets*

Réda DENAE 28

28

Performance Metrics: True Positive Rate, Recall, sensitivity

Recall, *True Positive Rate (TPR)*, or *Sensitivity* correspond to the proportion of positive data points that are correctly considered as positive, with respect to all positive data points.

Of all the really sick people, how many have we detect?

$$Recall = Sensitivity = True\ Positive\ Rate(TPR) = \frac{\text{True Positive}(TP)}{\text{Positive}(P)}$$

Recall is generally used in information retrieval applications, *recall* is the fraction of the relevant documents that are successfully retrieved

Réda DENAE 29

29

Performance Metrics: False Alarm Rate, False Positive Rate

False Alarm Rate (FAR) or *False Positive Rate (FPR)* is defined as the probability of falsely rejecting the null hypothesis.

$$False\ Alarm\ Rate\ (FAR) = False\ Positive\ Rate(FPR) = \frac{\text{False Positive}(FP)}{\text{Negative}(N)}$$

Réda DENAE 30

30

Performance Metrics: Missed Detection Rate, False Negative Rate

Missed Detection Rate (FAR) or False Negative Rate (FNR) is the proportion of the individuals with a known **negative** condition for which the test result is **positive**.

$$\text{Missed Detection Rate (MISS)} = \text{False Negative Rate (FNR)} = \frac{\text{False Negative (FN)}}{\text{Positive (P)}}$$

Réda DENAE 31

31

Performance Metrics: Specificity, True Negative Rate

Specificity or True Negative Rate (TNR) measures the proportion of really negatives examples identified as such

$$\text{Specificity} = \text{True Negative Rate (TNR)} = \frac{\text{True Negative (TN)}}{\text{Negative (N)}} = 1 - \text{FAR}$$

Réda DENAE 32

32

Performance Metrics: Precision and False Discovery Rate

- Precision* is the number of correct positive results divided by the number of positive results predicted by the classifier.
- False Discovery Rate (FDR)* is the expected proportion of False Positive (Type I errors)

$$\text{Precision} = \frac{\text{True Positive (TP)}}{(\text{TP} + \text{FP})} = 1 - \text{FDR}$$

$$\text{False Discovery Rate (FDR)} = \frac{\text{False Positive (FP)}}{(\text{TP} + \text{FP})} = 1 - \text{Precision}$$

Réda DENAE 33

33

Performance Metrics: Positive and Negative Predictive Value

Positive predictive value is the probability that subjects with a positive screening test truly are positive.

Negative predictive value is the probability that subjects with a negative screening test truly are negative.

$$\text{Positive Predictive Value (PPV)} = \frac{TP}{(TP + FP)}$$

$$\text{Negative Predictive Value (NPV)} = \frac{TN}{(TN + FN)}$$

Réda DENAE 34

34

Performance Metrics: F-Score, F-Measure

- F-Score:** $F\text{-Score} = \text{Precision} \times \text{Recall}$
- F-Measure:** Harmonic mean of Precision and Recall:

$$F_\beta = \frac{(1 + \beta^2) \times \text{Recall} \times \text{Precision}}{\beta^2 \times \text{Recall} + \text{Precision}}$$
- F₁-Measure:**

$$F_1 = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

Réda DENAE 35

35

Performance Curves

- Receiver Operating Characteristic (ROC) Curves
- Precision-Recall Curves
- Detection Error-Tradeoff (DET) Curves

Réda DENAE 36

36

Receiver Operating Characteristic (ROC) Curves

- **Receiver Operating Characteristic**
 - Used in signal detection theory: Tradeoffs in Hits vs. False alarms.
 - Medical diagnosis: Costs/tradeoffs in type-I, type-II errors
- **Data Mining**
 - Visualizing classifier performance
 - Comparing classifiers
 - Useful where:
 - Class distributions are unequal
 - Different misclassification costs

37

37

Receiver Operating Characteristic (ROC) Curves

Plot TPR vs. FPR as the classifier goes from “conservative” to “liberal”

38

38

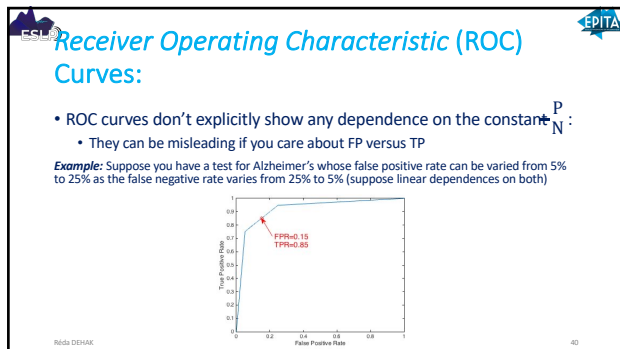
Receiver Operating Characteristic (ROC) Curves: Area Under the ROC Curve (AUC)

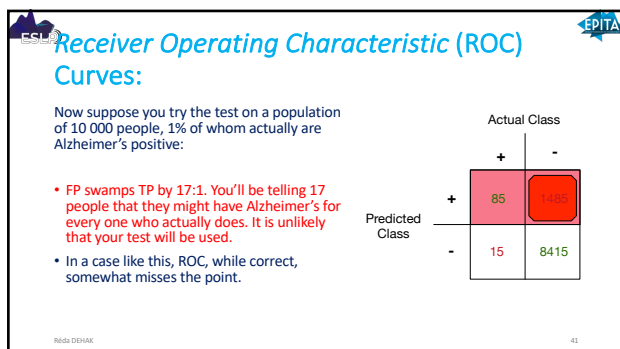
- The ROC curve is used to generate a summary statistic:
 - The **Area Under the ROC Curve (AUC, A' or c-statistic)**
 - The area between the ROC curve and the no-discriminate line
 - Gini coefficient:

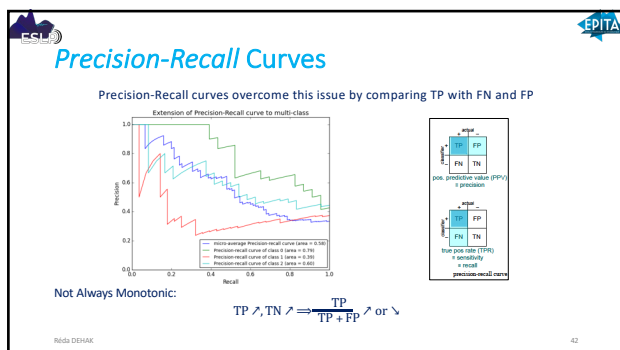
$$G_1 = 2 \text{ AUC} - 1$$
- Youden's J statistic: The intercept of the ROC curve with the line at $\pi/2$ to the no-discrimination line.

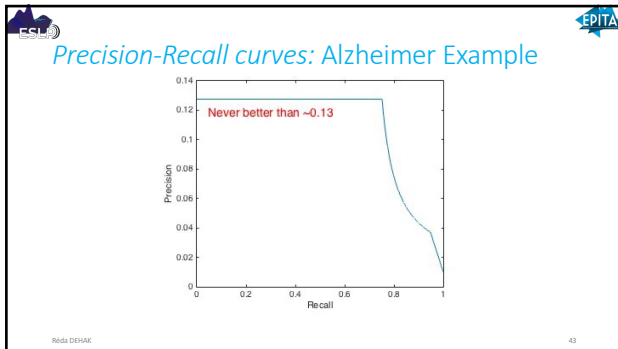
39

39

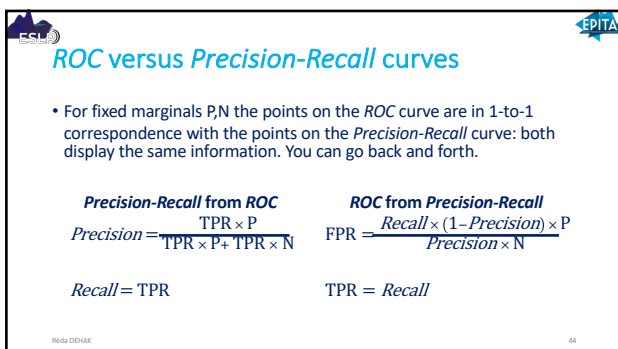




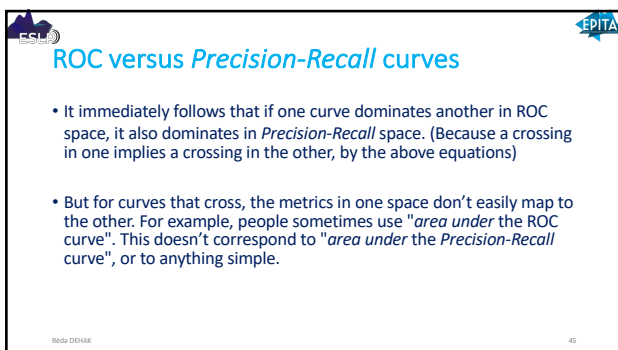




43



44

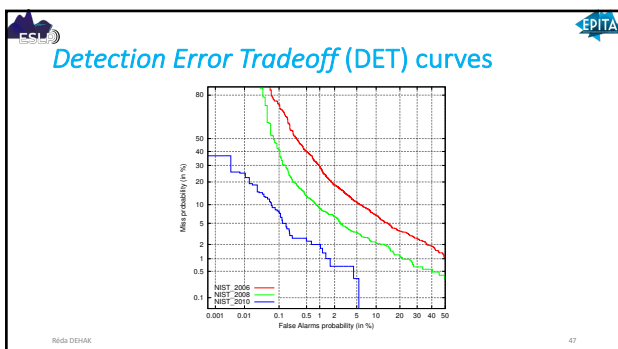


45

Detection Error Tradeoff (DET) curves

- An alternative to the ROC curve.
- Plot the Missed Detections (FNR) vs. the False Alarms (FPR).
- Non-linearly (logarithmic) transformed x- and y-axes (quantile function of the normal distribution)
- The DET plot is used extensively in the evaluation of biometric systems.

46





Detection Error Tradeoff (DET) curves

The DET curve is used to generate a summary statistic:

- **Equal Error Rate (EER):** The intercept of the DET curve with the line corresponding to $y = x$.
- **Detection Cost Function (DCF):** A weighted average of the missed detection and false alarm rates. The point on the DET Curve where such an average is minimized may be indicated (minDCF). If you have to provide a hard decision, the distance between the minDCF operating point and the operating point of this hard decision is an indication of how appropriately the system implementers chose the hard decision operating points to optimize the chosen cost function (Calibration).

48



Conclusion:

- Logistic Regression is a simple Linear Classifier
- Threshold must be fixed according to the cost (risk) function
- Confusion matrix is the best performance measure.

Réda DENAE49
