Linear Classification

Réda DEHAK
http://isml.dehak.org

1

Contents

- Classification
- Linear Classification
- Logistic Regression
- Performance metrics

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: $y \epsilon \{0, 1\}$
    - Email : Spam / Not Spam?
    - Online Transactions: Fraudulent (Yes/No)
    - Check Identity: Target / impostors
  - Multiclass classification:

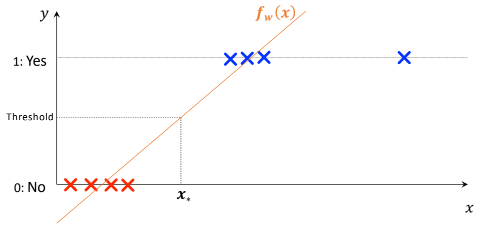
3

Classification

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

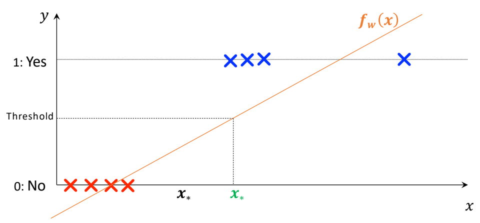
4

Classification as a Regression problem



5

Classification as a Regression problem



6

Logistic Regression

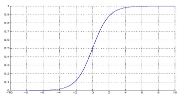
- Want $0 \leq f_W(x) \leq 1$:
  - Linear Regression:

$$f_W(x) = W^T x$$

  - Logistic Function:

$$h(z) = \frac{1}{1 + e^{-z}}$$

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



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

Réda DEHAK 7

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

Réda DEHAK 8

8

Logistic Regression

- $P(y = 1) \geq P(y = 0) \Leftrightarrow \frac{P(y=1)}{p(y=0)} \geq 1$

$$\Leftrightarrow log\left(\frac{P(y=1)}{p(y=0)}\right) \geq 0$$

$$\Leftrightarrow log\left(\frac{P(y=1)}{1-p(y=1)}\right) = logit(P(y = 1)) \geq 0$$

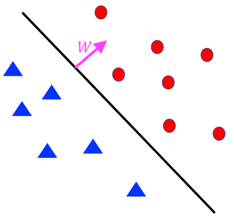
$$\Leftrightarrow logit(P(y = 1)) = \boldsymbol{W^T x}$$

Réda DEHAK 9

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



Réda DEHAK 10

10

Logistic Regression

- $P(y=1) \geq P(y=0) \Leftrightarrow \frac{P(y=1)}{p(y=0)} \geq 1$

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

Boundary = Hyperplane

Réda DEHAK 11

11

Linear Classification

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

Réda DEHAK 12

12

Training: Two Classes

$$\mathcal{X} = \{(x_{i,}\, y_i), i = 1 \ldots N\} \quad y_i \epsilon \{0,1\}$$
$$y_i \sim Bernoulli$$

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

Réda DEHAK 13

13

Training: Two Classes

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

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

Réda DEHAK 14

14

Gradient Descent for linear Regression

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

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

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

Réda DEHAK 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 DEHAK 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 DEHAK 17

17

Performance Metrics

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.

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

19

Performances: Missclassification Errors

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

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

20

Example:

Logistic Regression

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

SVM

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

Which is the Best Classifier?

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

Réda DEHAK 22

22

Threshold

Using costs values to adjust Threshold to minimize costs

Actual Class

Predicted Class	+	-
+	True Positives (TP) Cost = 0	False Positives (FP) Cost = 20
-	False Negatives (FN) Cost = 5	True Negatives (TN) Cost = 0

$Score \geq 0.6$

	+	-
+	40	20
-	10	30

$Score \geq 0.7$

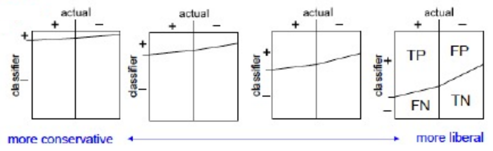
	+	-
+	30	10
-	20	40

Réda DEHAK 23

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!



Réda DEHAK 24

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.

25

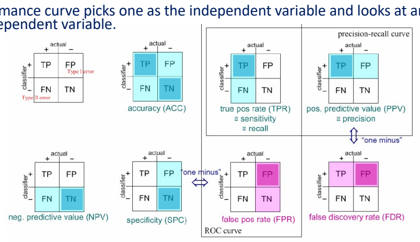
Performance Metrics

- 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

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.



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

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

Réda DEHAK 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?

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

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

Réda DEHAK 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.

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

Réda DEHAK 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.**

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

Réda DEHAK 31

31

Performance Metrics: *Specificity, True Negative Rate*

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

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

Réda DEHAK 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)

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

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

Réda DEHAK 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.

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

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

Réda DEHAK 34

34

Performance Metrics: *F-Score, F-Measure*

- ***F-Score:***

$$F\text{–}Score = Precision \times Recall$$

- ***F-Measure***: Harmonic mean of *Precision* and *Recall*:

$$F_\beta = \frac{(1+\beta^2) \times Recall \times Precision}{\beta^2 \times Recall + Precision}$$

- ***F_1-Measure:***

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

Réda DEHAK 35

35

Performance Curves

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

Réda DEHAK 36

36

Receiver Operating Characteristic (ROC) Curves

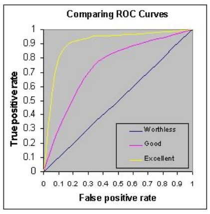
- ***Receiver Operating Characteristic***
  - Used in signal detection theory: Tradeoffs in Hits vs. False alarms.
  - Medical diagnosis: Costs/tradeoffs in type-I, type-II errors
- Data Mining
  - Visualizing classifier performance
  - Comparing classifiers
  - Useful where:
    - Class distributions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

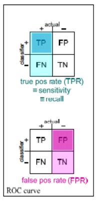
Réda DEHAK 37

37

Receiver Operating Characteristic (ROC) Curves

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





Réda DEHAK 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:

$$G_1 = 2\,\mathrm{AUC} - 1$$

- Youden's J statistic: The intercept of the ROC curve with the line at $\pi/2$ to the no-discrimination line.

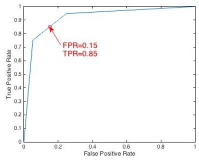
Réda DEHAK 39

39

Receiver Operating Characteristic (ROC) Curves:

- ROC curves don't explicitly show any dependence on the constant $\frac{P}{N}$:
  - They can be misleading if you care about FP versus TP

Example: Suppose you have a test for Alzheimer's whose false positive rate can be varied from 5% to 25% as the false negative rate varies from 25% to 5% (suppose linear dependences on both)



40

Receiver Operating Characteristic (ROC) Curves:

Now suppose you try the test on a population of 10 000 people, 1% of whom actually are Alzheimer's positive:

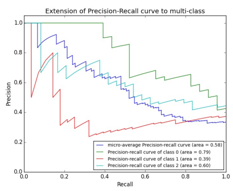
- FP swamps TP by 17:1. You'll be telling 17 people that they might have Alzheimer's for every one who actually does. It is unlikely that your test will be used.
- In a case like this, ROC, while correct, somewhat misses the point.

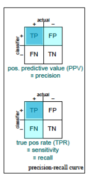
		Actual Class +	Actual Class -
Predicted Class	+	85	1485
	-	15	8415

41

Precision-Recall Curves

Precision-Recall curves overcome this issue by comparing TP with FN and FP



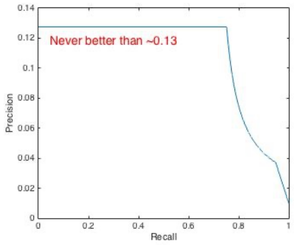


Not Always Monotonic:

$$TP \nearrow, TN \nearrow \Longrightarrow \frac{TP}{TP+FP} \nearrow \text{ or } \searrow$$

42

Precision-Recall curves: Alzheimer Example



43

ROC versus *Precision-Recall* curves

- For fixed marginals P,N the points on the *ROC* curve are in 1-to-1 correspondence with the points on the *Precision-Recall* curve: both display the same information. You can go back and forth.

Precision-Recall* from *ROC

$$Precision = \frac{\mathrm{TPR} \times \mathrm{P}}{\mathrm{TPR} \times \mathrm{P} + \mathrm{TPR} \times \mathrm{N}}$$

$$Recall = \mathrm{TPR}$$

ROC* from *Precision-Recall

$$\mathrm{FPR} = \frac{Recall \times (1 - Precision) \times \mathrm{P}}{Precision \times \mathrm{N}}$$

$$\mathrm{TPR} = Recall$$

44

ROC versus *Precision-Recall* curves

- It immediately follows that if one curve dominates another in ROC space, it also dominates in *Precision-Recall* space. (Because a crossing in one implies a crossing in the other, by the above equations)
- But for curves that cross, the metrics in one space don't easily map to the other. For example, people sometimes use "*area under* the ROC curve". This doesn't correspond to "*area under* the *Precision-Recall* curve", or to anything simple.

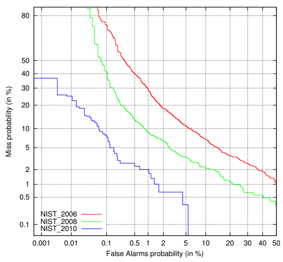
45

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



47

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 DEHAK 49

49