Machine Learning

By Ghazal Lalooha

Designing A Machine Learning System

Spam Detection

From: cheapsales@buystufffromme.com
To: ghazallalooha0gmail.com
Subject: Buy now!

Deal of the week! Buy now!
Rolex w4tchs - $100
Med1cine (any kind) - $50
Also low cost M0rgages available.

From: Sullivan Kettler
To: ghazallalooha0gmail.com
Subject: Christmas dates?

Hi Ghazal,
Was talking to Nicolas about plans for Xmas. When do you get off work. Meet Dec 22?
Sullivan.

Create a spam classifier

- Supervised learning:
  - x: email features
    - Selection of 100 words that represent spam or normal email, such as purchase, discount, deal, ...
  - y: spam (1) or non-spam (0)

- Note: In practice, n more common words (10,000 to 50,000) are usually used in the training set.

Create a spam classifier

- Question: How to create a classifier with low error rate?
  - Collect a lot of data
  - Development of advanced features based on email routing information
  - Development of advanced features based on the words used in the message body.
    - For example, whether words such as deal, deals and trader should be considered as one word or not.
  - Development of complex algorithms to detect intentional spelling mistakes!

Error Analysis

Recommended Approach

- Selection of learning algorithm and implementation:
  - Start with a simple algorithm that can be quickly implemented.
  - Implement it and test it on the validation set.
- Learning algorithm troubleshooting:
  - Draw learning curves to understand if you need more data, more features, less regularization, etc.
- Error analysis:
  - Check the data in the validation set where the algorithm got it wrong.
  - See if there is a common feature in this data that is causing the error.

Error Analysis

- Suppose there are 500 email samples in the validation set.
- The learning algorithm has misclassified 100 emails.
- Check out these 100 examples and categorize them by things like:
  - Email type: drug, advertising, password theft
  - The clues you think can help the algorithm in classifying these emails correctly.

Drug: 12
Advertising: 4
password theft: 53
Others: 31

Deliberate spelling mistakes: 5
Unusual routing: 16
Unusual use of punctuation: 32

The Importance of Quantitative Assessments and Numbers

- Question: Should words like deal, deals and trader be considered the same?
  - For this purpose, you can use software related to finding the roots of words. (like Porter Stemmer)
- Error analysis is not helpful in these cases and the only solution is to test the above idea in practice.
- In other words, we need to numerically evaluate the performance of the algorithm in both cases on the validation set and then make a decision based on the evaluation results.
  - No word rooting: 5% error
  - By rooting words: 3% error

Error Measurement for Unbalanced Classes

Example: Cancer diagnosis

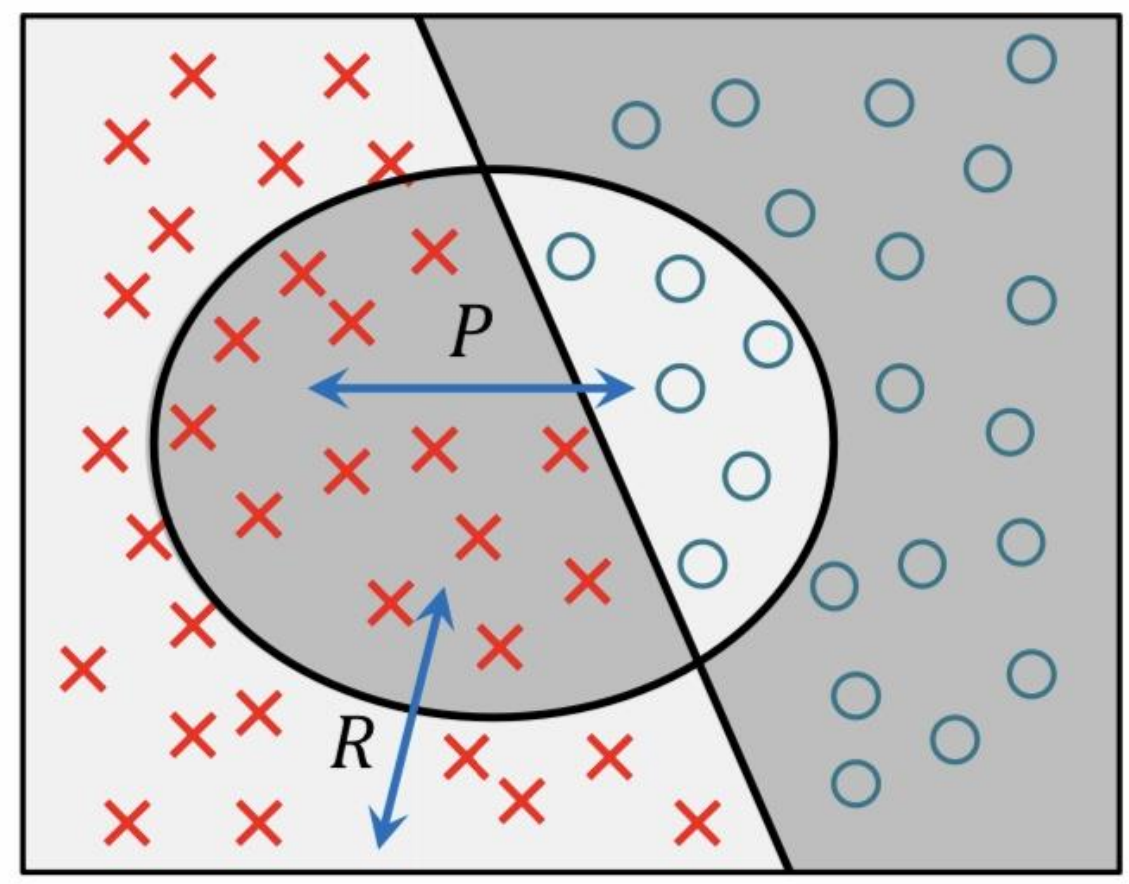
- Train a logistic regression model.
- Output: cancer (y = 1) otherwise (y = 0)
- Suppose the error of the trained model for the test set is 1%, (99% correct detection).
- Suppose in our data only 0.5% of patients actually have cancer.

```
function y = predictCancer(x)
     y = 0;   % just ignore x
end;
```

Error: 0.5 %

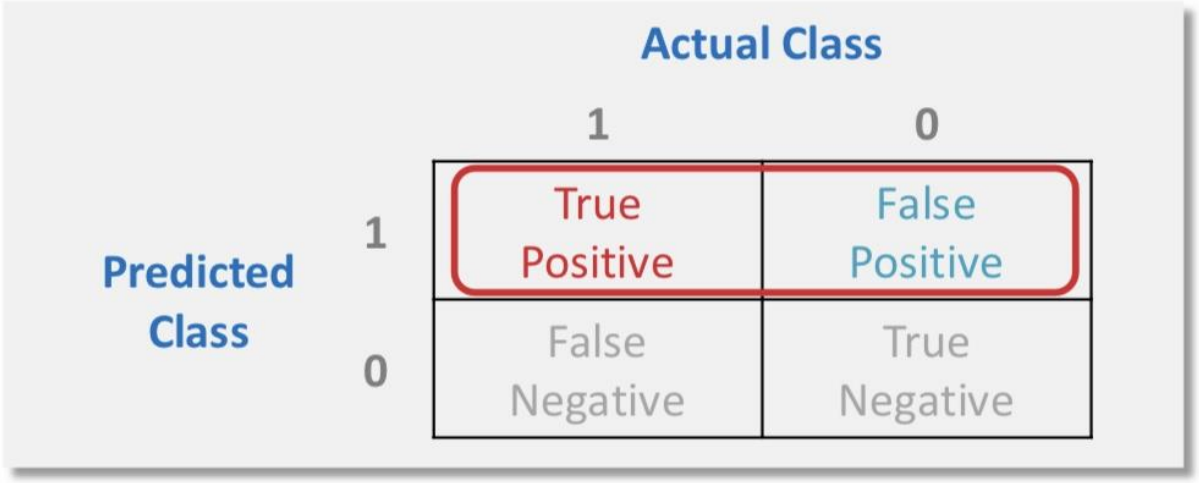
- Unbalanced class: A class in which the ratio of the number of positive samples to the number of negative samples (or vice versa) is very small (close to zero).

Precision and Recall



Precision

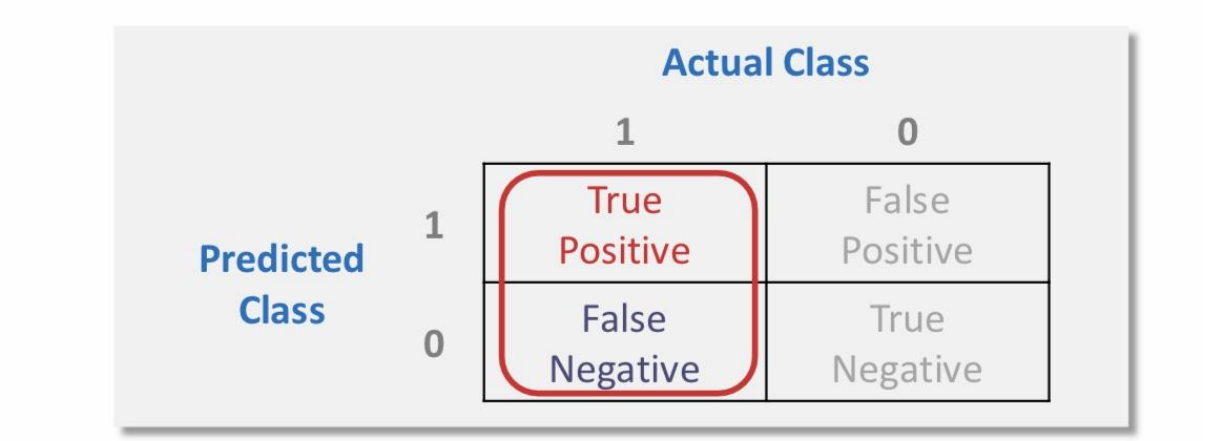
- Precision: The ratio of the number of samples that are correctly diagnosed as positive to the total number of samples that are diagnosed as positive.



$$\frac{\text{True Positive}}{\text{\# predicted Positive}} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

Recall

- Recall: The ratio of the number of samples that are correctly diagnosed as positive to the total number of samples that are truly positive.



$$\frac{\text{True Positive}}{\text{\# actual Positive}} = \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

Balance Between Precision and Recall

- Logistic regression:
  - Hypothesis: $0 \leq h_\theta(x) \leq 1$
  - Forecast:
    - y = 1: if $h_\theta(x) \geq 0.5$
    - y = 0: if $h_\theta(x) < 0.5$

$$\text{precision} = \frac{\text{true positives}}{\text{no. of predicted positive}}$$

$$\text{recall} = \frac{\text{true positives}}{\text{no. of actual positive}}$$

- In order to increase the reliability factor, it can be predicted as follows:
  - y = 1: if $h_\theta(x) \geq 0.9$
  - y = 0: if $h_\theta(x) < 0.9$

precision increase

recall decrease

Balance Between Precision and Recall

- Logistic regression:
  - Hypothesis: $0 \leq h_\theta(x) \leq 1$
  - Forecast:
    - y = 1: if $h_\theta(x) \geq 0.5$
    - y = 0: if $h_\theta(x) < 0.5$

$$\text{precision} = \frac{\text{true positives}}{\text{no. of predicted positive}}$$

$$\text{recall} = \frac{\text{true positives}}{\text{no. of actual positive}}$$

- In order to increase the reliability factor, it can be predicted as follows:
  - y = 1: if $h_\theta(x) \geq 0.3$
  - y = 0: if $h_\theta(x) < 0.3$

precision decrease

recall increase

F Score

- Question: How can you compare the precision rate and recall rate of different algorithms?

	Recall rate	Precision rate
Algorithm one	0.5	0.4
Algorithm two	0.7	0.1
Algorithm three	0.02	1.0

- F Score:
  - If P = 0 or R = 0, then F Score is equal to zero.
  - If P = 1 or R = 1, then F Score is equal to one.

$$2\frac{P \cdot R}{P + R}$$

F Score

- Question: How can you compare the precision rate and recall rate of different algorithms?

	Recall rate	Precision rate	F Score
Algorithm one	0.5	0.4	0.444
Algorithm two	0.7	0.1	0.175
Algorithm three	0.02	1.0	0.039

- F Score:
  - If P = 0 or R = 0, then F Score is equal to zero.
  - If P = 1 or R = 1, then F Score is equal to one.

$$2\frac{P \cdot R}{P + R}$$

Data for Machine Learning

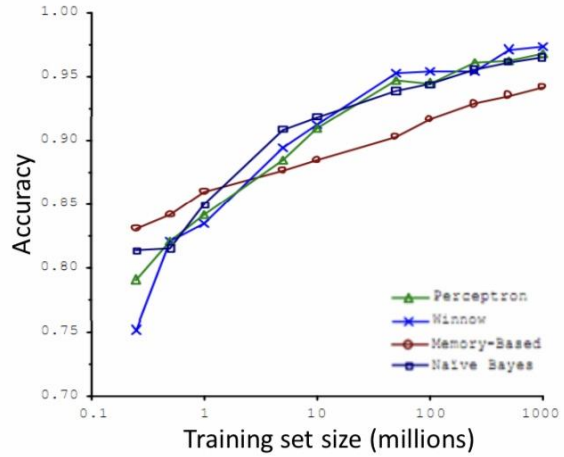
Designing a Learning System with High Accuracy

- Problem: recognizing similar words (Banko and Brill, 2001)

  {to, too, two}, {then, than}

  For breakfast, I ate_____ eggs.

- Algorithms:
  - perceptron (logistic regression)
  - Vino
  - Memory based
  - Biz classifier



"The winner is not the one with the best algorithm. The winner is the one with the most data."

Rationale for Big Data

- Note: Having more data is only useful when the feature vector x contains enough information to estimate the output y.
  - Example 1: Fill in the blanks with the given words (yes)
  - Example 2: Estimating the house price only by having its size (no)

- A useful experiment: given the input x, can an expert predict the value of y with high confidence?

Rationale for Big Data

- Using a powerful learning algorithm with many parameters:
  - Linear or logistic regression with a very large number of features.
  - Neural network with a very large number of hidden units
  - Low training set error (1)
- Using a very large training set (reducing the risk of overfitting)
  - The error of the training set is almost equal to the error of the test set (2).
- Result (1) and (2): Low test set error. (high generalizability)