Algorithms for Machine Learning

Chiranjib Bhattacharyya

Dept of CSA, IISc
chibha@chalmers.se

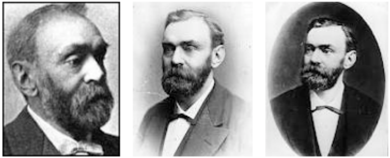
January 17, 2012

Agenda

- Introduction to classification
- Bayes Classifier

Who is the person?

Images of one person



Who is the person?

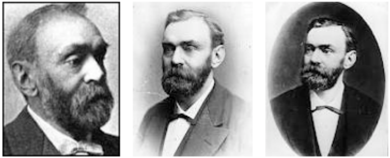
Images of one person



Is he the same person?

Who is the person?

Images of one person



Is he the same person?
easy

Who is the person?

Images of one person

Is he the same person?

Who is the person?

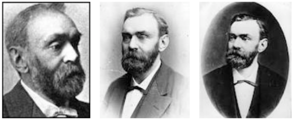
Images of one person



Is he the same person?
not so easy

Who is the person?

Images of one person

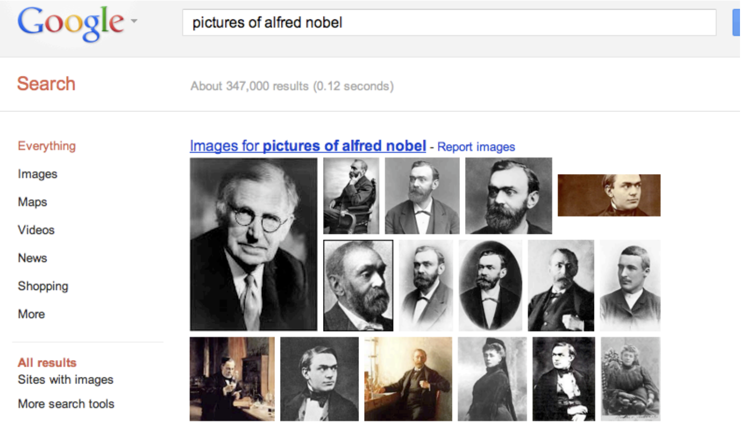


Is he the same person?
not so easy

But who is he?

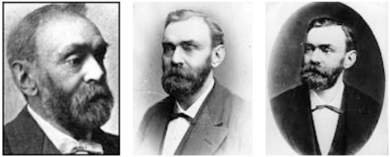
ALFRED NOBEL

Introduction to Classification



Lots of scope for improvement.

The classification problem setup



Alfred Nobel

Bertha Von Suttner

Objective

From these images create a function, classifier, which can automatically recognize images of Nobel and Suttner

The steps

Step 1 Create representation from the Image, sometimes called a feature map.

Step 2 From a training set and a feature map create a classifier

Step 3 Evaluate the goodness of the classifier

We will be concerned about Step 2 and Step 3.

The classification problem setup

- Let $(\mathbf{X}, Y) \sim P$ where P is a Distribution and

  $$D_m = \{(\mathbf{X}_i, Y_i) |\ i.i.d\ \mathbf{X}_i, Y_i \sim P, i = 1, \ldots, m\}$$

  is a random sample
- Probability of misclassification

  $$R(f) = P(f(\mathbf{X}) \neq Y)$$

Finding the best classifier

- Suppose $P(Y = y|X = x)$ was high then it is very likely that that x has the label y.
- Define $\eta(x) = P(Y = 1|X = x)$, posterior probability computed from Bayes rule from Class-conditional densities $P(X = x|Y = y)$
- For 2 classes, $f^*(x) = \text{sign}(2\eta(x) - 1)$ is the Bayes classifier.

Finding the best classifier

Objective should be to choose f such that

$$min_f R(f)$$

Theorem

Let f be any other classifier and f^* be Bayes Classifier

$$R(f) \geq R(f^*)$$

A very important result

Bayes Classifier has the least error rate. $R(f^*)$ is called the Bayes error-rate.

- Review Maximum Likelihood estimation
- Try to construct Bayes Classifier

Naive Bayes Classifier

- Assume that the features are independent
- works well for many problems, specially on *text classification*

Spam Emails

Search Results | Delete

Subject: Your Email Address Has Made You A Millionaire!!!
From: "The Awards Committee" <candexinfo@bellnet.ca>
Date: Tue, September 9, 2008 8:11 am
To: winners@euromillions.org
Priority: Normal
Options: View Full Header | View Printable Version | Download this as a file

You Have Been Selected As Winner Of A Cash Prize

Of One Million Euro (€1,000,000.00 EURO). For More

Information Contact Mr. Donald Wong, Email:donaldwong.anzbnk@live.com

Regards,

Ms. Roxanne Presley.

Spam Emails

Current Folder: INBOX

Compose Addresses Folders Options Search Help

Search Results | Delete

Subject: AWARD NOTIFICATION
From: "UK NATIONAL LOTTERY" <info@winners.com>
Date: Wed, March 25, 2009 6:32 pm
Priority: Normal
Options: View Full Header | View Printable Version | Download this as a file

```
P.O.Box 1010
Liverpool
L70 1NL
United Kingdom.
Ref: XYL /26510460037/05
Batch: 24/00319/IPD
Ticket Number : 56475600545188                         AWARD NOTIFICATION

This is to inform you that you have been  selected for a cash prize
of£1,500,000.00 pounds held  on  the 25th MARCH,2009 in London UK.
The selection process was carried out through random selection
Ourcomputerized email selection system(ess)
from a database of over 250,000  email Addresses drawn from which you were
selected.

You are to contact the fiduciary claims department by your personal
information with e-mail Giving Below;
```

Naive Bayes Classifier: Bernoulli model

Create a feature list where each feature is on/off. Denote the feature map $x = [f_1, \ldots, f_d]^\top$
$P(X = x | Y = y) = \prod_{i=1}^{d} P(F_i = f_i | Y = y)$

$$p_{1i} = P(F_i = 1 | Y = 1) \; p_{2i} = P(F_i = 1 | Y = 2)$$

Bayes Classifier: Output the class with the higher score

$$score_1(x) = \sum_i (f_i log p_{1i} + (1 - f_i) log(1 - p_{1i}))$$

similarly $score_2(x)$

Naive Bayes: Bernoulli

Source: Introduction to Information Retrieval. (Manning, Raghavan, Schutze)

TRAINBERNOULLINB($\mathbb{C}, \mathbb{D}$)
1 $V \leftarrow$ EXTRACTVOCABULARY($\mathbb{D}$)
2 $N \leftarrow$ COUNTDOCS($\mathbb{D}$)
3 **for each** $c \in \mathbb{C}$
4 **do** $N_c \leftarrow$ COUNTDOCSINCLASS($\mathbb{D}, c$)
5 $\quad$ $prior[c] \leftarrow N_c / N$
6 $\quad$ **for each** $t \in V$
7 $\quad$ **do** $N_{ct} \leftarrow$ COUNTDOCSINCLASSCONTAININGTERM($\mathbb{D}, c, t$)
8 $\quad\quad$ $condprob[t][c] \leftarrow (N_{ct} + 1)/(N_c + 2)$
9 **return** $V, prior, condprob$

APPLYBERNOULLINB($\mathbb{C}, V, prior, condprob, d$)
1 $V_d \leftarrow$ EXTRACTTERMSFROMDOC(V, d)
2 **for each** $c \in \mathbb{C}$
3 **do** $score[c] \leftarrow \log prior[c]$
4 $\quad$ **for each** $t \in V$
5 $\quad$ **do if** $t \in V_d$
6 $\quad\quad$ **then** $score[c] \mathrel{+}= \log condprob[t][c]$
7 $\quad\quad$ **else** $score[c] \mathrel{+}= \log(1 - condprob[t][c])$
8 **return** $\arg\max_{c \in \mathbb{C}} score[c]$

► **Figure 13.3** NB algorithm (Bernoulli model): Training and testing. The add-one smoothing in Line 8 (top) is in analogy to Equation (13.7) with $B = 2$.

Discriminant functions

Bayes Classifier

$$h(x) = \text{sign}\left(\sum_{i=1}^{d} f_i\theta_i - b\right)$$

$\theta_i = \log \frac{p_{1i}(1-p_{2i})}{(1-p_{1i})p_{2i}}$

$h(x)$ is sometimes called Discriminant functions

Gaussian class conditional distributions

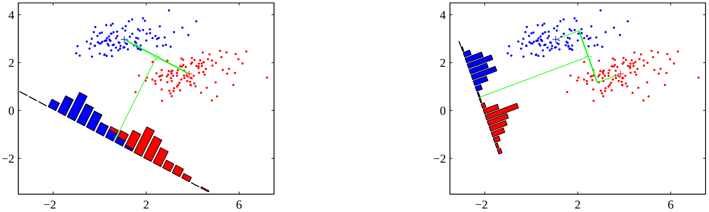
Let the class conditional distributions be $N(\mu_1, \Sigma)$ and $N(\mu_2, \Sigma)$. The Bayes classifier is given by

$$h(x) = \text{sign}(w^\top x - b)$$

$w = \Sigma^{-1}(\mu_1 - \mu_2)$

Fisher Discriminant

Source: Pattern Recognition and Machine Learning (Chris Bishop)



Fisher Discriminant

Let (μ_1, Σ_1) be the mean and covariance of class 1 and (μ_2, Σ_2) be the mean and covariance of class 2.

$$J(w) = max_w \frac{\left(w^\top(\mu_1 - \mu_2)\right)^2}{w^\top S w}$$

$w = S^{-1}(\mu_1 - \mu_2) \;\; S = \Sigma_1 + \Sigma_2$