Modelling and Classifying Random Phenomena

Colin Bellinger

School of Computer Science
Carleton University
Ottawa, Canada

cbelling@scs.carleton.ca

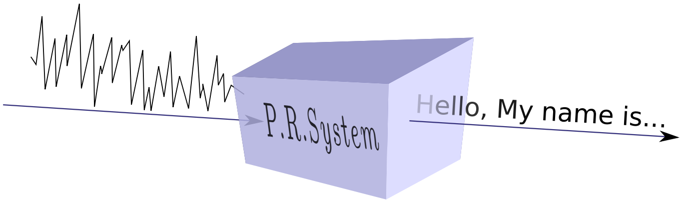
12 March, 2010

Overview

- Pattern Recognition
  - nomenclature
  - existing approaches
- Data generation
  - CTBT
  - dispersion modelling and simulation
- Preliminary results

Classification - Overview

- Automated identification of classes:
  - digits: 0 - 10
  - vehicles: car/truck
  - typist: a specific user
  - disease: cancer/not cancer

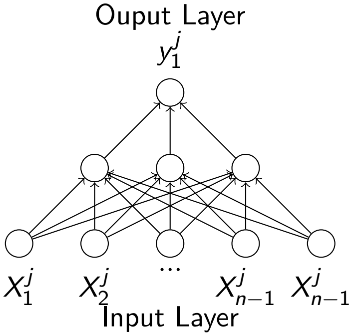


Classification - Binary versus One-class learning

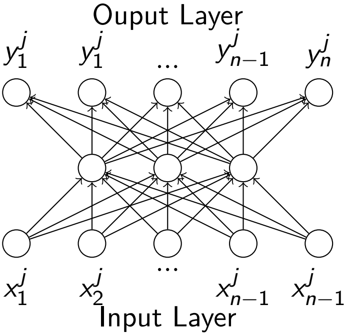
- **Binary**: concept learning process utilizes examples from both classes in the binary classification task
  - *Given labelled examples of two classes, define a function to identify new unlabelled examples*
- **One-class**: concept learning process utilizes examples from single class in the binary classification task
  - *Given labelled examples of a single class, define a function to identify new unlabelled examples of that target class.*

Classification - Classifiers

- **Artificial Neural Network**: MLP, Autoassociator
  - MLP: output is the actual classification
  - Autossociator: aims to reproduce (recognise) input at output layer



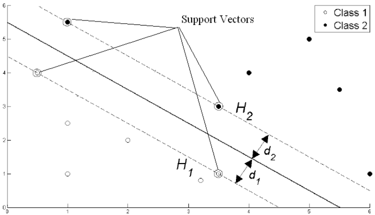
Multi-Layer Perceptron

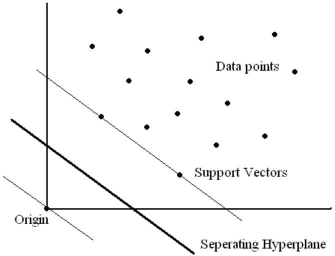


Autoassociator

Classification - Classifiers

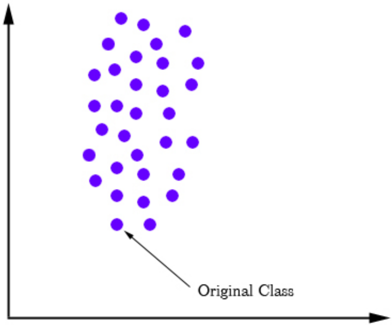
- **SVM**: binary and one-class
- Define a hyperplane which maximizes the gap.





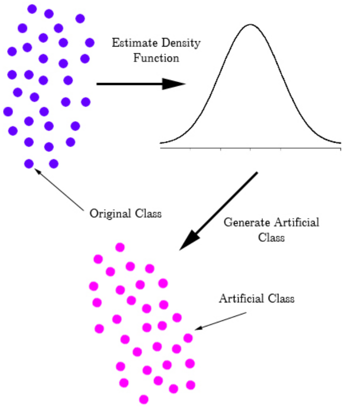
Classification - Classifiers

- Combined Density and Class Probability Estimator
- Step 1: Examine the data points of the positive class.



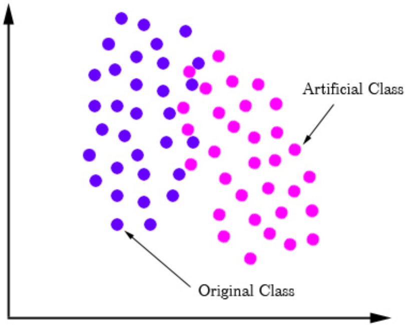
Classification - Classifiers

- Step 2: Determine the reference distribution, such as normal or multi-variate normal distribution, of the positive class.



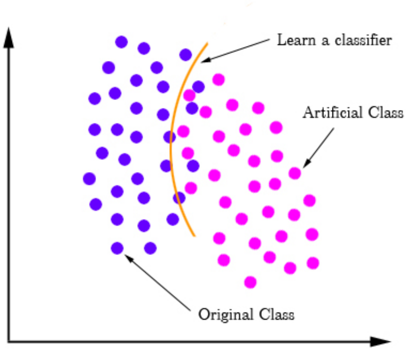
Classification - Classifiers

- Step 3: Then use our knowledge of the distribution to generate points around the positive class.



Classification - Classifiers

- Step 4: Apply a standard binary classi[U+FB01]er.



Classification - Classifier comparison

CDCPE versus bagged decision tree (BDT)

- AUC averaged over 15 UCI datasets
  - CDCPE: 0.843
  - BDT: 0.940

CDCPE versus libSVM

- FAR AND IPR averaged over 15 UCI datasets
  - CDCPE: 0.147, 0.157
  - libSVM: 0.113, 0.331

- Discrimination-based approaches are general more robust.
- However, one-class learners can be competitive.

The Comprehensive Test Ban Treaty (CTBT)

The CTBT is a United Nations treaty which will **bans all nuclear explosions in the environment** when it enters into force.
http://www.ctbto.org/

Simulating Dispersion

Objective:

- Generate a dataset containing a series of radioxenon measurements
- Receptor-specific datasets contain feature vectors of:
  - cumulative quantity of radioxenon measured over 12 or 24 hours
  - class label (background or explosion)

Modelling the Atmosphere

- Lagrangian particle models:
  - mathematically disperse pollutants via Markovian process
  - each step depends on current atmospheric conditions
- Gradient transfer models:
  - gradient parameters define diffusion
  - wind speed defines down-wind advection
- Gaussian models:
  - distribution of pollutant assumes a Gaussian form
  - wind speed defines down-wind advection

Modelling the Atmosphere

Gaussian puff model:

- sol'n to Fickian diffusion equation
- models diffusion from an instaneous point source of emission strength Q
- assume mean concentration of dispersing pollutant forms a Gaussian distribution

$$\overline{\chi}(x,y,z,t) = \frac{Q}{(4\pi t)^{\frac{3}{2}}(K_x K_y K_z)^{\frac{1}{2}}} exp\left[-\left(\frac{(x-\overline{u}t)^2}{4K_x t} + \frac{y^2}{4K_y t} + \frac{z^2}{4K_z t}\right)\right]$$

Modelling the Atmosphere

Gaussian plume model:

- models diffusion from a continuous point source (Q), emitted from an elevated industrial stack
- infinite number of puffs superimposed on each other
- mathematically speaking, integrate with respect to time
- as a matter of convenience, diffusion along x-axis is ignored

$$\chi(x, y, z, t) = \frac{Q}{2\pi\sigma_y\sigma_z\overline{u}} exp\left(- \left(\frac{y^2}{2\sigma_y^2} + \frac{z^2}{2\sigma_z^2} \right) \right)$$

Modelling the Atmosphere

Gaussian plume model:

- account for reflection at surface

$$\chi(x, y, z, t) = \frac{Q}{2\pi\sigma_y\sigma_z\overline{u}} exp\left(-\frac{y^2}{2\sigma_y^2}\right)$$
$$\left[exp\left(-\frac{(z-H)^2}{2\sigma_z^2}\right) + exp\left(-\frac{(z+H)^2}{2\sigma_z^2}\right)\right]$$

where:

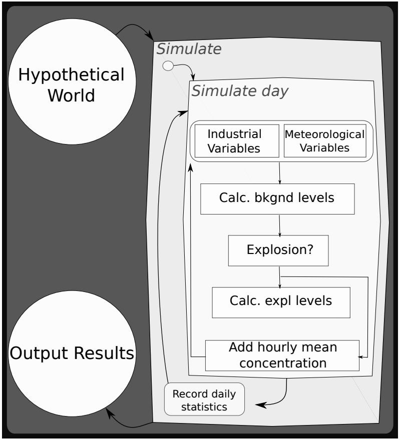
h is the height of the plumes centreline

- in much the same way, reflection at an inversion layer can be accounted for

Simulating Dispersion

1. Define hypothetical world
2. Simulation for j=1:n days
   (i) For each day, simulate i=1:24 hours
      - generate Gaussian random variables about the means
      - calculate background radioxenon levels
      - if explosion, added expls levels to bkgnd levels
      - add hourly mean to cumulative daily count

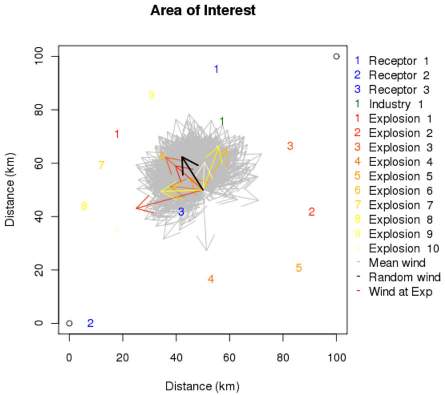
   (ii) Record daily value
3. Output dataset



Simulating Dispersion

Sample map:

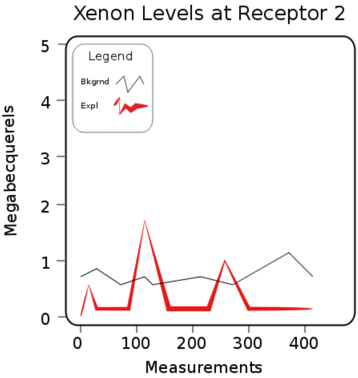
- 1 industrial emitter (green)
- 3 receptors (blue)
- 10 explosions (heat colours)



Simulating Dispersion

plotted results for receptor 2:

- background data in black
- explosions in red
- generally up-wind from industry $\longrightarrow$ low background levels
- two main peaks (expl 4 and expl 5)



Preliminary Results

Classifier	Class	TPR	FPR	AUC
MLP	target	0.845	0	0.896
	outlier	1	0.155	0.896
J48	target	0.997	0	0.990
	outlier	1	0.003	0.990
IBK	target	0.995	0	0.998
	outlier	1	0.005	0.998
NB	target	0.039	0.1	0.754
	outlier	0.990	0.964	0.754
CDCPE	target	0.902	0.013	0.650
	outlier	0.087	0.098	0.650

Conclusion

- Examined strategies for modelling atmospheric dispersion
- In the spirit of the CTBT
  - applied a Gaussian assumption to model the dispersion of radioxenon
  - generate background noise and random phenomena
- Utilized Weka to classify the preliminary dataset