Novelty detection

Slides by Refael Chickvashvili and downloadable from

http://www.cs.huji.ac.il/course/2004/learns/

P. Agius – L14, Spring 2008

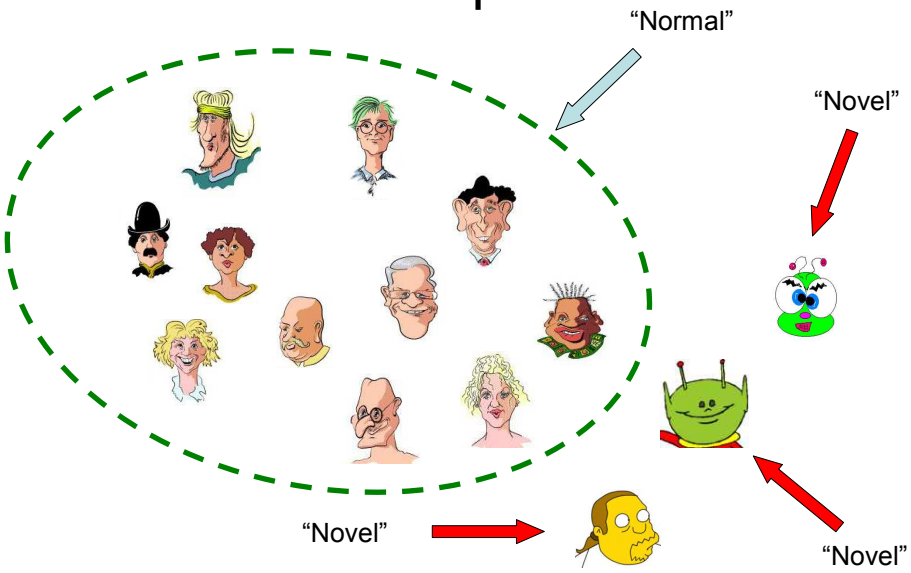
Outline

- Introduction
- Quantile Estimation
- OCSVM – Theory
- OCSVM – Application to Jet Engines

Novelty Detection is

- An unsupervised learning problem (data unlabeled)
- About the identification of new or unknown data or signal that a machine learning system is not aware of during training

Example 1



So what seems to be the problem?

It's a [illegible] problem.
"Normal vs. "Novel"

Wrong!

The Problem is

That "All positive examples are alike but each negative example is negative in its own way".

Example 2

- Suppose we want to build a classifier that recognizes web pages about “pickup sticks”.
- How can we collect a training data?
  - We can surf the web and pretty easily assemble a sample to be our collection of positive examples.
- What about negative examples ?
  - The negative examples are… the rest of the web. That is ~(“pickup sticks web page”)
- So the negative examples come from an unknown # of negative classes.

Applications

- Many exist
  - Intrusion detection
  - Fraud detection
  - Fault detection
  - Robotics
  - Medical diagnosis
  - E-Commerce
  - And more…

Possible Approaches

- Density Estimation:
  - Estimate a *density* based on training data
  - Threshold the estimated density for test points
- Quantile Estimation:
  - Estimate a *quantile* of the distribution underlying the training data: for a fixed constant $\alpha \in (0,1]$, attempt to find a small set S such that $\Pr(x \in S) = \alpha$
  - Check whether test points are inside or outside S

Quantile Estimation (QE)

- A quantile function with respect to (P, λ, H) is defined as :

$$U(\mu) = \inf\{\lambda(C) \mid P(C) \geq \mu, C \in H\}$$

- $0 < \mu \leq 1$
- H- a class of measurable subsets of X
- λ - a real valued function. $\lambda : H \to \Re$
- $C_\lambda(\mu)$ denotes the $C \in H$ that attains the infimum

Quantile Estimation (QE)

- The empirical quantile function is defined as above where P is the empirical distribution:

$$P_{emp}^{m}(C) = \frac{1}{m}\sum_{i=1}^{m} I_C(x_i)$$

- $C_{\lambda}^{m}(\mu)$ - denotes the $C \in H$ that attains the infimum on the training set.

- Thus the goal is to estimate $C_{\lambda}(\mu)$ through $C_{\lambda}^{m}(\mu)$

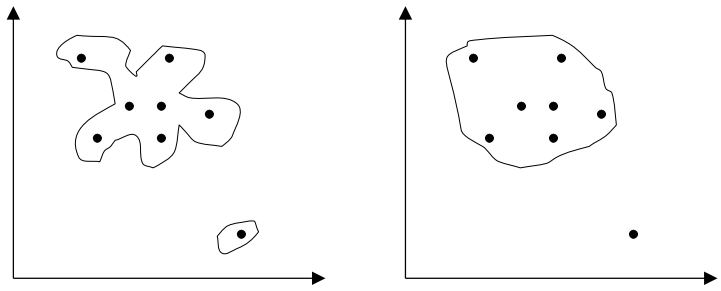
Quantile Estimation

- Choosing Intelligently H and λ is important

- On one hand large class H → many small sets that contain a fraction μ of the training examples.

- On the other hand, if we allowed just *any* set, the chosen set could consist of only the training points → poor generalization

Complex vs. Simple



$P^m_{emp}(C) = 1$ $\quad$ $P^m_{emp}(C) < 1$

Support Vector Method for Novelty Detection

Bernhard Schölkof, Robert Williams, Alex Smola, John Shawe-Taylor, John Platt

Problem Formulation

- Suppose we are given a training ample drawn from an underlying distribution P
- We want to a estimate a "simple" subset $S \subset X$ such that for a test point x drawn from the distribution P, $\Pr(x \notin S) = \nu, \nu \in (0,1]$
- We approach the problem by trying to estimate a function f which is positive on S and negative on the complement

The SV Approach to QE

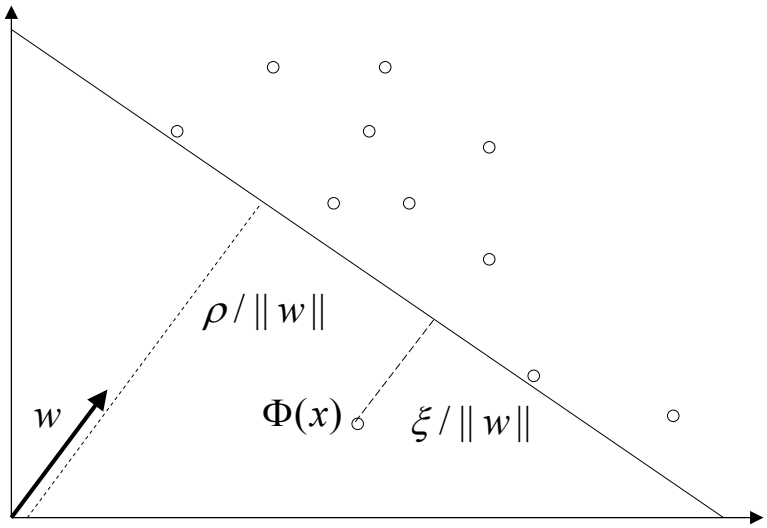
- The class H is defined as the set of half-spaces in a feature space F (via kernel k)
- Here we define $\lambda(C_w) = \| w \|^2$, where $C_w = \{x \mid f_w(x) \geq \rho\}$
- (w, ρ) are respectively a weight vector and an offset parameterizing a hyperplane in F

"Hey, Just a second"

If we use hyperplanes & offsets, doesn't it mean we separate the "positive" sample? But, separate from what?

From the Origin

OCSVM



OCSVM

Serves as a penalizer like "C" in the 2-class svm (recall that $0 < \nu \leq 1$)

$$\min_{w \in F, \xi_i \in R, \rho \in R} \frac{1}{2} \| w \| + \frac{1}{\nu m} \sum_i \xi_i - \rho$$

subject to $\langle w, \Phi(x) \rangle \geq \rho - \xi_i$

Notice that no "y"s are incorporated in the constraint since there are no labels

OCSVM

- The decision is therefore:

$$f(x) = \operatorname{sgn}(\langle w, \Phi(x) \rangle - \rho)$$

- Since the slack variables ξ_i are penalized in the objective function, we can expect that if w and ρ solve the problem then f will equal 1 for most example in the training set, while $\| w \|$ still stays small

OCSVM

Using multipliers $\alpha_i, \beta_i \geq 0$ we get the Lagrangian,

$$L(\mathbf{w}, \boldsymbol{\xi}, \rho, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2} \| \mathbf{w} \|^2 + \frac{1}{\nu m} \sum_i \xi_i - \rho - \\ - \sum_i \alpha_i (\langle \mathbf{w}, \Phi(\mathbf{x}) \rangle - \rho + \xi_i) - \sum_i \beta_i \xi_i$$

OCSVM

Setting the derivatives of L w.r.t $\mathbf{w}, \boldsymbol{\xi}, \rho$ to 0 yields:

1) $\mathbf{w} = \sum_i \alpha_i \Phi(x_i)$

2) $\alpha_i = \frac{1}{\nu m} - \beta_i \leq \frac{1}{\nu m}$, $\sum_i \alpha_i = 1$

OCSVM

Eq. 1 transforms $f(x)$ into a kernel expansion:

$$f(\mathbf{x}) = \operatorname{sgn}\left(\sum_i \alpha_i k(\mathbf{x}_i, \mathbf{x}) - \rho \right)$$

The offset ρ can be recovered by exploiting that for any $0 < \alpha_i < \frac{1}{\nu m}$ the corresponding pattern $\mathbf{x}_i$ satisfies:

$$\rho = \langle \mathbf{w}, \Phi(\mathbf{x}_i) \rangle = \sum_j \alpha_j k(\mathbf{x}_j, \mathbf{x}_i)$$

ν- Property

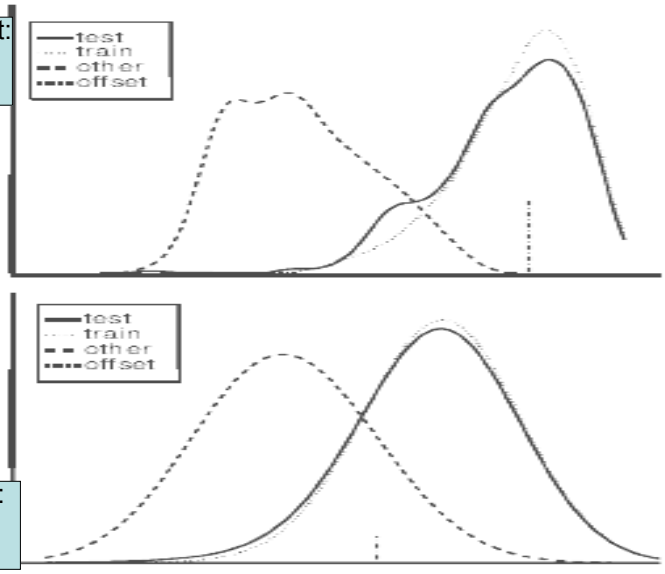
- Assume the solution of the primal problem satisfies $\rho \neq 0$ The following statements hold:
  - ν is an upper bound on the fraction outliers.
  - ν is a lower bound on the fraction SVs
  - With probability 1, asymptotically, ν equals both the fraction of SVs and the fraction of outliers. (under certain conditions of P(x) and the kernel)

X axis – svm magnitude

y axis – frequency

Results – USPS (“0”)

For ν = 50%, we get:
50% SVs
49% Outliers



For ν = 5%, we get:
6% SVs
4% Outliers

OCSM - Shortcomings

- Implicitly assumes that the “negative” data lies around the origin.
- Ignores completely “negative” data even if such data partially exist.

Support Vector Novelty Detection Applied to Jet Engine Vibration Spectra

Paul Hyton, Bernhard Schölkof,
Lionel Tarassenko, Paul Anuzis

Intro.

- Jet engines have pass-off tests before they can be delivered to the customer.
- Through vibration tests an engine's "vibration signature" can be extracted
- While normal vibration signatures are common, we may be short of abnormal signatures.
- Or even worse, the engine under test may show up a type of abnormality which has never been seen before.

Feature Selection

- A vibration gauges are attached to the engine's case
- The engine under test is slowly accelerated from idle to full speed and decelerated back to idle
- The vibration signal is then recorded
- The final feature is calculated over a weighted average of the vibration for 10 different speed ranges
- Thus yielding a 10-D vector

Algorithm

- Slightly more general than the regular OCSVM
- In addition to the "normal" data points $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_m\}$ we take into account some abnormal points $\mathbf{Z} = \{\mathbf{z}_1, ..., \mathbf{z}_t\}$
- Rather than separating from the origin we separate from the mean of $\mathbf{Z}$

Primal Form

$$\min_{w\in F,\xi_i\in R,\rho\in R} \frac{1}{2}\|\mathbf{w}\| + \frac{1}{\nu m}\sum_i \xi_i - \rho$$

subject to $\langle \mathbf{w}, (\mathbf{x}_i - \frac{1}{t}\sum_n \mathbf{z})\rangle \geq \rho - \xi_i$

$$\xi_i \geq 0$$

and the decision function is

$$f(\mathbf{x}) = \mathrm{sgn}(\langle \mathbf{w}, (\mathbf{x} - \frac{1}{t}\sum_n \mathbf{z}_n)\rangle - \rho)$$

Dual Form

$$\min_{\boldsymbol{\alpha}\in R^m} \frac{1}{2}\sum_{ij} \alpha_i\alpha_j (k(\mathbf{x}_i,\mathbf{x}_j) - q - q_j - q_i)$$

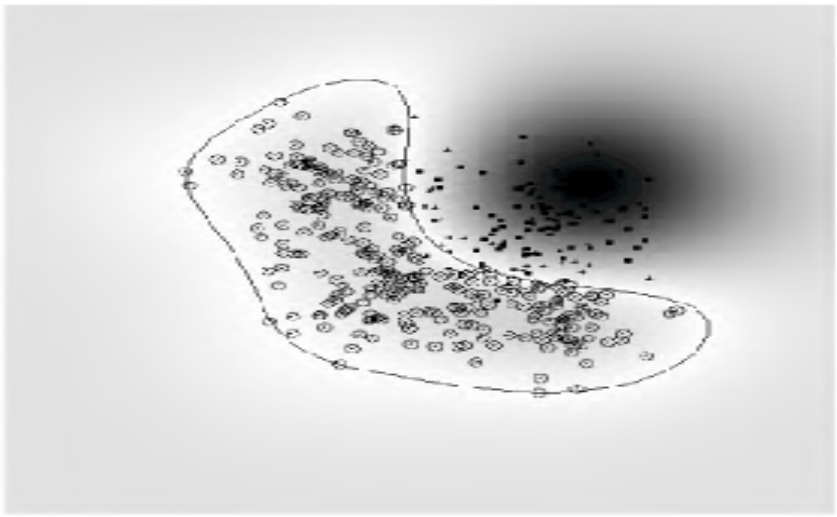
where

$$q = \frac{1}{t^2}\sum_{np} k(\mathbf{z}_n,\mathbf{z}_p) \quad \text{and} \quad q_j = \frac{1}{t}\sum_n k(\mathbf{x}_j,\mathbf{z}_n)$$

subject to

$$0 \leq \alpha_i \leq \frac{1}{\nu m} \quad , \quad \sum_i \alpha_i = 1$$

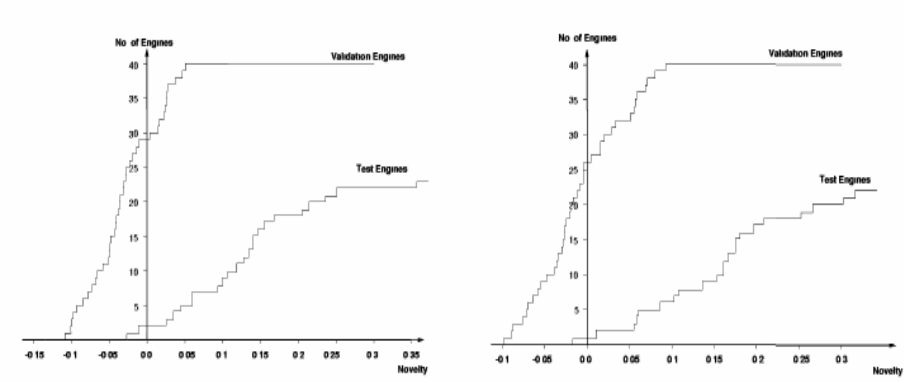
2D Toy Example



Training Data

- 99 Normal Engines were used as training data
- 40 Normal Engines were used as validation data
- 23 Abnormal Engines used as test data

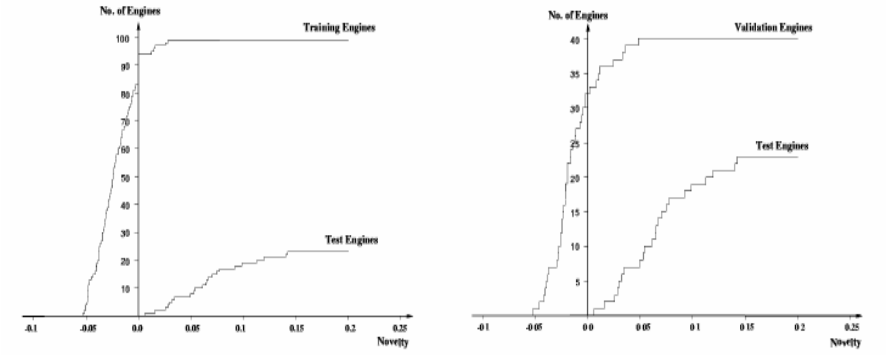
Standard OCSM Results



(a) $\nu = 0.1$ (b) $\nu = 0.2$

Figure 2: Cumulative novelty distributions for two different values of ν. The curves show that there is a slight overlap in the data; For $\nu = 0.1$, there are 11 validation engines ove the SVM decision boundary and 2 test engines inside the boundary.

Modified OCSM Results



(a) ($\nu = 0.1$) (b)

Figure 3: Cumulative novelty distributions showing the variation of novelty with number of engines for (a) the training data versus the test data (each test engine omitted from the training phase in turn to compute its novelty) and (b) the validation data versus the test data.