Machine Learning 10-701 Mar 16, 2015
http://alex.smola.org/teaching/10-701-15 due on Mar 24, 2015
Carnegie Mellon University

Homework 7

START HERE: Instructions

- The homework is due at 9:00am on Mar 24, 2015. Anything that is received after that time will not be considered.
- Answers to every theory questions will be also submitted electronically on Autolab (PDF: Latex or handwritten and scanned). Make sure you prepare the answers to each question separately.
- Collaboration on solving the homework is allowed (after you have thought about the problems on your own). However, when you do collaborate, you should list your collaborators! You might also have gotten some inspiration from resources (books or online etc...). This might be OK only after you have tried to solve the problem, and couldn't. In such a case, you should cite your resources.
- If you do collaborate with someone or use a book or website, you are expected to write up your solution independently. That is, close the book and all of your notes before starting to write up your solution.
- Latex source of this homework: http://alex.smola.org/teaching/10-701-15/homework/hw7_latex.zip

1 Novelty Detection [Manzil; 60 points]

In this problem we will look into the task of novelty detection. Novelty detection is an unsupervised task where one is interested in flagging a small fraction of the input dataset as atypical or novel. It can be viewed as a special case of the quantile estimation task, where we are interested in estimating a simple set $\mathcal{C}$ such that $\Pr(x \in \mathcal{C}) \geq \mu$ for some $\mu \in [0,1]$. One way to measure simplicity is to use the volume of the set. Formally, if $|\mathcal{C}|$ denotes the volume of a set, then the novelty detection or quantile estimation task is to estimate arginf$\{|\mathcal{C}| s.t. \mathcal{C}) \geq \mu\}$. The importance of anomaly detection is due to the fact that novelty in data translate to significant (and often critical) actionable information in a wide variety of application domains. For example, some atypical traffic pattern in a computer network could mean that a hacked computer is sending out sensitive data to an unauthorized destination. An anomalous MRI image may indicate presence of malignant tumors. Anomalies in credit card transaction data could indicate credit card or identity theft or anomalous readings from a space craft sensor could signify a fault in some component of the space craft.

1.1 Simple Minimum Enclosing Balls

Consider a data set $X = \{x_1, ..., x_n\}$ of n observations from the same distribution described by d features. We want to learn a function $f(\cdot)$ which can classify if the new observation so different from the others that we can doubt it is regular? Or on the contrary, is it so similar to the other that we cannot distinguish it from the original observations? This is the question addressed by the novelty detection tools and methods.

A simple idea to find novel points would be to learn a ball which packs most of the points, or in general a rough, close frontier delimiting contour which packs most of the points. Now, if the observations lay within the frontier-delimited subspace, then are considered as typical observations. Otherwise, if they lie outside the frontier, we can say that they are abnormal with a given confidence in our assessment.

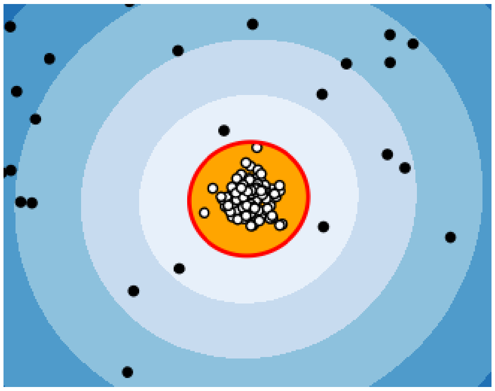


Figure 1: A simple dataset with outliers

Question 1

To begin with suppose, we want to find the minimum enclosing ball for the given data of n points $x_1, ..., x_n \in \mathbb{R}^d$. That is, find the centre $z \in \mathbb{R}^d$ and the smallest radius $R \geq 0$ satisfying

$$\forall i : \quad \|x_i - z\|^2 \leq R^2 \tag{1}$$

Your goal is to state this in the form of a convex optimization problem and explain why this is a convex problem.

Let us now relax the constraint to pack all the points in the ball. How will we modify the optimization problem by introducing slack variables?
Hint: It is similar to SVM with slack variables.

Then the points left out of the ball by solving the optimization problem can be considered as the novel points and the fraction of novel points is controlled by the regularization constant on the slack variables.

1.2 Minimum Enclosing Balls in Feature Space

While such single minimum enclosing balls may seem rather restrictive. For example consider the dataset shown in figure 2, it has two different blobs where data is concentrated. In such cases the idea of finding a simple enclosing ball may not be sufficient. Remember that the kernel trick can be used to map data into a high-dimensional space and simple decision boundaries in the mapped space correspond to highly non-linear decision boundaries in the input space.
Thus, we map to a high dimensional feature space to handle such cases. Suppose we use a RBF Mercer Kernel $k(x, x') = \langle \Phi(x), \Phi(x') \rangle$ and instead try to find a minimum enclosing ball in the feature space. By RBF (*radial basis function*) Mercer Kernel, we refer to kernels which only depend on $x - x'$, i.e. $k(x, x') = k(x - x')$ and $k(x, x)$ is constant.

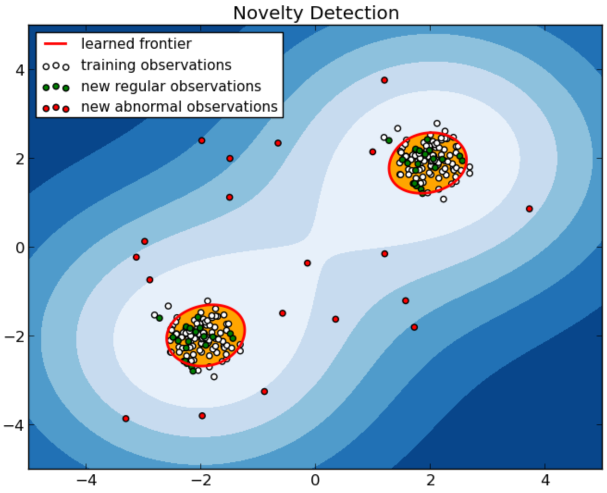


Figure 2: A complex dataset with outliers

Question 2

Given n points $x_1, ..., x_n \in \mathbb{R}^d$, we want to find enclosing ball of the data in the feature space. That is, find the centre $z \in \mathbb{R}^f$ and the smallest radius $R \geq 0$ satisfying

$$\forall i : \quad \|\Phi(x_i) - z\|^2 \leq R^2 + \xi_i, \tag{2}$$

where $\xi_i \geq 0$. Also derive the dual of the optimization problem.

1.3 Finding Maximum Distance Separating Hyperplane

Now let us take a detour and focus on one of the method covered in the class for novelty detection, namely the method prescribed in unit 4 Linear Methods, slide 191. It is built on an idea which appears to be bizarre at first sight, i.e. finding a hyperplane in the feature, given by $f(x) = \langle w, \Phi(x)\rangle = \rho$ that has maximum distance from origin yet is still closer to the origin than the observations. Then we classify novelty for a point x, by determining which side of the hyperplane $\langle w, \Phi(x)\rangle = \rho$ does the point falls on. In this problem we will try to demystify this idea.

Question 3

Given n points $x_1, ..., x_n \in \mathbb{R}^d$ lying in a orthant, we want to find a hyperplane $\langle w, x\rangle = \rho$ that has maximum distance from origin and separating most of the points from the origin. That is, find the normal vector $w \in \mathbb{R}^d$ to the plane satisfying

$$\forall i : \langle w, x_i\rangle \geq \rho - \xi_i, \tag{3}$$

where $\xi_i \geq 0$. Your goal is to state this in the form of a convex optimization problem and explain why this is a convex problem. What if the points were mapped to feature space through $\Phi(\cdot)$, ?

Answer

To separate the data set from the origin, we solve the following quadratic program:

$$\begin{aligned} \min_{w,\rho,\xi_i} \quad & \frac{1}{2}||w||^2 - \rho + C\sum_{i=1}^{n} \xi_i \\ \text{subject to} \quad & \langle w, x_i\rangle \geq \rho - \xi_i \;\; \text{where } \xi_i \geq 0 \;\forall i \in [n] \end{aligned} \tag{4}$$

It is convex because:

- Quadratic objective with positive definite matrix and linear terms
- Linear constraints

In case of mapping to a feature space, we have only minimal change:

$$\begin{aligned} \min_{w,\rho,\xi_i} \quad & \frac{1}{2}||w||^2 - \rho + C\sum_{i=1}^{n} \xi_i \\ \text{subject to} \quad & \langle w, \Phi(x_i)\rangle \geq \rho - \xi_i \;\; \text{where } \xi_i \geq 0 \;\forall i \in [n] \end{aligned} \tag{5}$$

1.4 Establishing the Equivalence

Now we show that actually both of these problems are actually the same through some nice geometric properties of Reproducing Kernel Hilbert Space.

Question 4

Let $k(\cdot, \cdot)$ be a RBF Mercer kernel, i.e. $k(x, x') = k(x - x')$. Then show that in the feature space, any dataset of n points $x_1, ..., x_n \in \mathbb{R}^d$ will lie on surface of a hypersphere in the same orthant.

So no matter how badly the data is distributed in original space upon transforming to the higher dimensional feature space using a RBF Mercer Kernel, the mapped points are always confined to an octant. Therefore, in feature space finding the smallest sphere (containing the mapping of the points) really amounts to finding the smallest segment of the sphere that the data live on. The segment, however, can be found in a straightforward way by simply intersecting the data sphere with a hyperplane. The hyperplane with maximum margin of separation to the origin will cut off the smallest segment, thus giving rise to the single class ν-SVM formulation based on finding the maximum distance separating hyperplane.

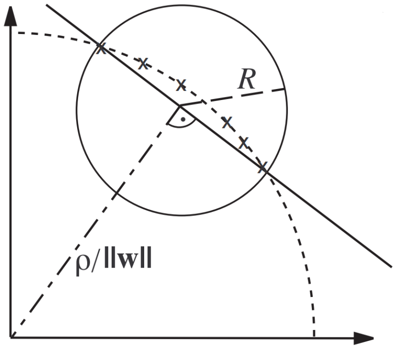


Figure 3: Finding enclosing ball and max distance plane are equivalent

Question 5

Derive the dual of the problem (5) and thus verify that indeed finding maximum distance separating hyperplane in the feature space is indeed equivalent to finding the minimum enclosing ball.

Hopefully this provides an intuition why single class SVM tries to find Maximum Distance Separating Hyperplace in anomaly detection. With proper parametrization of the regularization constant we can also obtain famous ν property which can be used to control fraction of novel points as can be found in proposition 1 of http://alex.smola.org/papers/2000/SchWilSmoShaetal00.pdf.

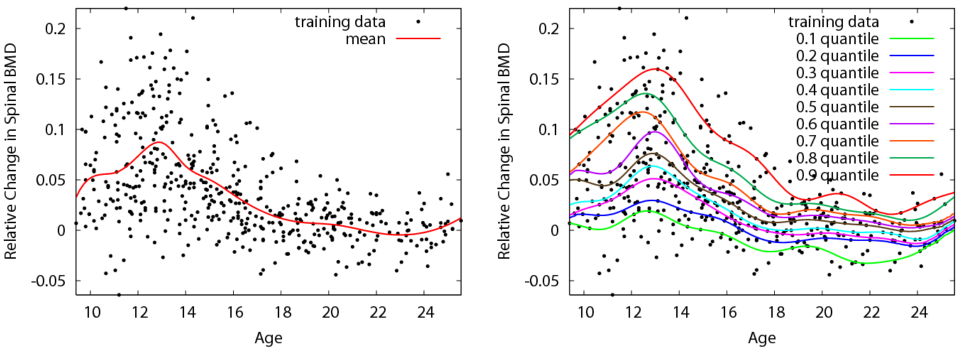
Machine Learning 10-701 Mar 16, 2015
http://alex.smola.org/teaching/10-701-15 due on Mar 24, 2015
Carnegie Mellon University

Homework 7

2 Quantile Regression [Jay-Yoon; 30 points]

In this problem we consider a very similar, but a more general task to novelty detection in the setting of regression. Generally, in a typical regression setting, the prediction of y at a given x is through conditional mean. Au contraire, the desired estimate of $y|x$ is not always the conditional mean, as sometimes one may want to obtain a good estimate that satisfies the property that a proportion, τ , of $y|x$, will be below the estimate (somewhat similar to novelty detection definition). For $\tau = 0.5$ this is an estimate of the median. What might be called median regression, is subsumed under the term quantile regression (QR) where the definition of the quantile is given below.

Definition 1 *(Quantile) Denote by $y \in \mathbb{R}$ a random variable and let $\tau \in (0, 1)$. Then the τ-quantile of y, denoted by μ_τ is given by the infimum over μ for which $\Pr\{y \le \mu\} = \tau$. Likewise, the conditional quantile $\mu_\tau(x)$ for a pair of random variables $(x, y) \in \mathcal{X} \times \mathbb{R}$ is defined as the function$\mu_\tau : \mathcal{X} \longrightarrow \mathbb{R}$ for which pointwise μ_τ is the infimum over μ for which $\Pr\{y \le \mu|x\} = \tau$.*



(a) Conditional mean analysis (b) Conditional quantile analysis

Figure 4: Comparsion between mean and quantile regression

There exist a large area of problems where estimating quantile, rather than mean, could be helpful and following are few of the motivating examples:

- A device manufacturer may wish to know what are the 10%and 90% quantiles for some feature of the production process, so as to tailor the process to cover 80% of the devices produced.
- For risk management and regulatory reporting purposes, a bank may need to estimate a lower bound on the changes in the value of its portfolio which will hold with high probability.
- A pediatrician requires a growth chart for children given their age and perhaps even medical background, to help determine whether medical interventions are required, e.g. while monitoring the progress of a premature infant.

Machine Learning 10-701 Mar 16, 2015
http://alex.smola.org/teaching/10-701-15 due on Mar 24, 2015
Carnegie Mellon University

Homework 7

2.1 Quantile Estimator

Let $Y = y_1, \ldots, y_m \subset \mathbb{R}$, $\mu \in \mathbb{R}$ and let $\tau \in (0,1)$. Observe that minizing ℓ_1-loss function ,i.e. $\sum_{i=1}^{m} |y_i - \mu|$, over μ for a an estimator yields the median. To obtain the τ-quantile in a similar fashion, we define qunatile the "pinball" loss $l_\tau(\xi)$.

$$l_\tau(\xi) = \begin{cases} \tau\xi & \text{if } \xi \geq 0 \\ (\tau - 1)\xi & \text{if } \xi < 0, \end{cases} \tag{6}$$

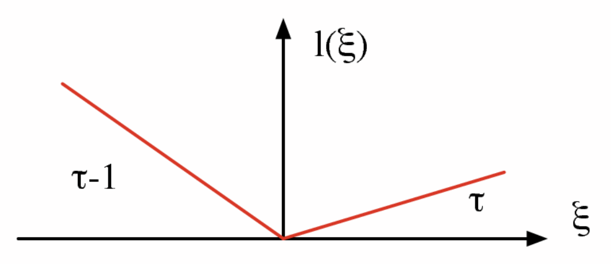


Figure 5: Outline of "pinball" loss $l_\tau(\xi)$.

Question 6

Prove the following properties about the minizier μ_τ of $\sum_{i=1}^{m} l_\tau(y_i - \mu_\tau)$:

1. The number of terms, m_-, with $y_i < \mu_\tau$ is bounded from above by τm.
2. The number of terms, m_+, with $y_i > \mu_\tau$ is bounded from above by $(1-\tau)m$.
3. For $m \to \infty$, the fraction $\frac{m_-}{m}$, converges to τ if $\Pr(y)$ does not contain discrete components.

These properties about the the loss function $l_\tau(\xi)$ indicates that it is a good candidate for QR. We will explore its use in QR in the next sub-problem.

Homework 7

2.2 Quantile Risk Optimization

Based on $l_\tau(\xi)$, we define the expected quantile risk as

$$R[f] \equiv \mathbf{E}_{p(x,y)}[l_\tau(y - f(x))].$$

For $f : \mathcal{X} \longrightarrow \mathbb{R}$, the minimizer of $R[f]$ is the quantile $\mu_\tau(x)$. As we are unaware of he true distribution $p(x, y)$, we define empirical risk function with additional regularizer function:

$$\begin{aligned} R_{reg}[f] &\equiv \frac{1}{m}\sum_{i=1}^{m} l_\tau(y_i - f(x_i)) + \frac{\lambda}{2}||w||_2^2 \\ \text{where} \qquad & f(x_i) = \langle \Phi(x_i), w\rangle + b \end{aligned} \tag{7}$$

where $f(x_i) = \langle \Phi(x_i), w\rangle + b$ denotes general linear models we dealt in class with some mapping function $\Phi(x_i)$. Using this empirical risk, the τ-quantile regression function can be found by solving the following optimization problem

$$\min_{w,b} \frac{1}{m}\sum_{i=1}^{m} l_\tau(y_i - f(x_i)) + \frac{\lambda}{2}||w||_2^2. \tag{8}$$

Question 7

Show that the minimization problem (8) is equivaelent to the following minimization problem (9):

$$\begin{aligned} \min_{w,b,\xi_i,\xi_i^{(*)}} \quad & \frac{1}{m}\sum_{i=1}^{m}\left(\tau\xi_i + (1-\tau)\xi_i^{(*)}\right) + \frac{\lambda}{2}||w||^2 \\ \text{subject to} \quad & y_i - \langle \Phi(x_i), w\rangle - b \le \xi_i \ \text{ and} \\ & \langle \Phi(x_i), w\rangle + b - y_i \le \xi_i^{(*)} \text{ where } \xi_i, \xi_i^{(*)} \ge 0 \end{aligned} \tag{9}$$

Also derive the dual of the problem (9). Use $K(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j)\rangle$ for notation.

If you solved problem 1 and 2 correctly, you should find the optimization problem is very similar. Also note the similarity of quantile regression to that of an ϵ insensitive support vector regression (ϵ-SVR) estimator. The key difference between the two estimation problems is that in (ϵ-SVR) we have an additional $\epsilon\|\alpha\|_1$ penalty in the objective function. This ensures that observations with deviations from the estimate, i.e. with $|y_i f(x_i)| < \epsilon$ do not appear in the support vector expansion. Moreover the upper and lower constraints on the Lagrange multipliers α_i are matched. This means that we balance excess in both directions. The latter is useful for a regression estimator. In our case, however, we obtain an estimate which penalizes loss unevenly, depending on whether $f(x)$ exceeds y or vice versa. This is exactly what we want from a quantile estimator: by this procedure errors in one direction have a larger influence than those in the converse direction, which leads to the shifted estimate we expect from QR. The practical advantage of the QR method discussed in this problem is that it can be solved directly with standard quadratic programming code rather than using pivoting, as is needed in ϵ-SVR.