

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

6.867 MACHINE LEARNING, FALL 2009

Problem Set 2

Due Date: Friday, October 9 (at 5pm)

NOTE: you will not receive a graded problem set in return for submitting one. We have adopted a new grading scheme for problem sets. We will help you understand and work through the problems. You are therefore strongly encouraged to clear up any misunderstandings prior to the deadline. By submitting your solutions, you indicate a) that you do understand the problem and b) that you have worked through the solution. Your submission will be recorded and you will receive no other grade from the submission. By completing a problem set, you demonstrate an achievement. In total, problem sets count 30% of the course grade.

Electronic submission is required! A submission form will be made available at the course website. You should have received the password required for submission via email.

Please see the course website for help on finding and using Matlab. Note that you will need the Optimization Toolbox, which is included in the MIT distribution.

Problem 1 Radial basis kernel

We can write the radial basis kernel in the following form:

$$K(\underline{x}, \underline{x}') = \exp\left\{-\frac{1}{2\sigma^2}\|\underline{x} - \underline{x}'\|^2\right\}, \quad (1)$$

where σ is a width parameter specifying how quickly the kernel vanishes as the points move further away from each other. This kernel has some remarkable properties. Indeed, we can perfectly separate *any* finite set of *distinct* training points. Moreover, this result holds for any positive finite value of σ . While the kernel width does not affect whether we'll be able to perfectly separate the training points, it does affect generalization performance. We will try to understand both of these issues a bit better.

(a) Let's proceed in stages. To make things easier we are going to prove a bit stronger result than we need to. In particular, we'll show that

$$\text{minimize } \frac{1}{2}\|\underline{\theta}\|^2 \quad \text{subject to } y_i \underline{\theta} \cdot \underline{\phi}(\underline{x}_i) = 1, \quad i = 1, \dots, n \quad (2)$$

has a solution regardless of how we set the ± 1 training labels y_i . You should convince yourself first that this is consistent with our goal. Here $\underline{\phi}(\underline{x}_i)$ is the feature vector (function actually) corresponding to the radial basis kernel. Our formulation here is a bit non-standard for two reasons. We try to find a solution where *all* the points are support vectors. This is not possible for all valid kernels but makes it easier to prove the result. We also omit the bias term since it is not needed for the result.

Introduce Lagrange multipliers for the constraints similarly to finding the SVM solution (see also the tutorial on Lagrange multipliers distributed along with the lecture slides) and show the form that the solution $\hat{\underline{\theta}}$ has to take. You can assume that $\underline{\theta}$ and $\underline{\phi}(\underline{x}_i)$ are finite vectors for the purposes of these calculations. Note that the Lagrange multipliers here are no longer constrained to be positive. Since you are trying to satisfy equality constraints, the Lagrange multipliers can take any real value.

We are after $\hat{\underline{\theta}}$ as a function of the Lagrange multipliers. (this should not involve lengthy calculations).

(b) Put the resulting solution back into the classification (margin) constraints and express the result in terms of a linear combination of the radial basis kernels.

(c) Indicate briefly how we can use the following Michelli theorem to show that any n by n kernel matrix $K_{ij} = K(\underline{x}_i, \underline{x}_j)$ for $i, j = 1, \dots, n$ is invertible.

Theorem: If $\rho(t)$ is monotonic function in $t \in [0, \infty)$, then the matrix $\rho_{ij} = \rho(\|\underline{x}_i - \underline{x}_j\|)$ is invertible for any distinct set of points $\underline{x}_i, i = 1, \dots, n$.

(d) Based on the above results put together the argument to show that we can indeed find a solution where all the points are support vectors.

(e) Of course, the fact that we can in principle separate any set of training examples does not mean that our classifier does well (on the contrary). So, why do we use the radial basis kernel? The reason has to do with margin that we can attain by varying σ . Note that the effect of varying σ on the margin is not simple rescaling of the feature vectors. Indeed, for the radial basis kernel we have

$$\phi(\underline{x}) \cdot \phi(\underline{x}) = K(\underline{x}, \underline{x}) = 1 \quad (3)$$

Let's begin by setting σ to a very small positive value. What is the margin that we attain in response to any n distinct training points?

(f) Provide a 1-dimensional example to show how the margin can be larger than the answer to part e). You are free to set σ and the points so as to highlight how they might "contribute to each other's margin".

Problem 2 Anomaly detection

One way to do anomaly detection is to find a tight enclosing ball of the feature vectors. Mathematically, given $\underline{x}_1, \dots, \underline{x}_n$, and a feature mapping $\phi(\underline{x})$, we solve

$$\min_{R, \underline{\theta}} R^2 \text{ subject to } \|\underline{\theta} - \phi(\underline{x}_i)\|^2 \leq R^2, \text{ for all } i = 1, \dots, n \quad (4)$$

so that $\underline{\theta}$ represents the center of the enclosing ball of radius R . In class, starting from a max-margin linear separator through origin, we derived an anomaly detection method by separating the feature vectors from the origin, formulated as

$$\min_{\rho, \underline{\theta}} \frac{1}{2} \|\underline{\theta}\|^2 - \rho \text{ subject to } \underline{\theta} \cdot \phi(\underline{x}_i) \geq \rho \text{ for all } i = 1, \dots, n \quad (5)$$

(a) Show that these are identical problems whenever $\|\phi(\underline{x}_i)\| = c$ for $i = 1, \dots, n$ and some constant c . In other words, the solution $\underline{\theta}^*$ obtained from either formulation should be the same.

(b) Finding the minimum enclosing ball in the feature space is also advantageous to solve in the dual. If we include slack variables, and set the slack penalty so as to omit a fraction ν of the positive training examples, we end up solving the following dual problem with respect to $\alpha = \{\alpha_1, \dots, \alpha_n\}$

$$\max_{\alpha} \sum_{i=1}^n \alpha_i K(\underline{x}_i, \underline{x}_i) - \sum_{i,j=1}^n \alpha_i \alpha_j K(\underline{x}_i, \underline{x}_j) \text{ subject to } 0 \leq \alpha_i \leq \frac{1}{\nu n}, \quad i = 1, \dots, n, \quad \sum_{i=1}^n \alpha_i = 1 \quad (6)$$

where $\underline{\theta}(\alpha^*) = \sum_{i=1}^n \alpha_i^* \phi(\underline{x}_i)$. Notice that the center of the ball lies in the convex hull of the feature vectors. What is the expression for R^* in terms of α^* and the kernel? What if you were given $\underline{\theta}(\alpha^*)$?

(c) Please modify the skeleton MATLAB code distributed with the problem set to solve for the minimum enclosing ball in the dual. You will need to modify / fill-in steps in `build_meb.m` and `meb_discriminant_function.m`.

(d) Load the data in `data.mat` to study the effect of different feature mappings (kernels) on the resulting minimum enclosing ball. We have provided you with the following kernel functions to try

$$K(\underline{x}, \underline{x}') = \underline{x} \cdot \underline{x}' \text{ (linear kernel K1)}$$

$$K(\underline{x}, \underline{x}') = (\underline{x} \cdot \underline{x}') + (\underline{x} \cdot \underline{x}')^2 \text{ (quadratic kernel K2)}$$

$$K(\underline{x}, \underline{x}') = \exp(-1/2\|\underline{x} - \underline{x}'\|^2) \text{ (radial basis kernel Kr)}$$

You can learn the minimum enclosing ball by calling `meb = meb_build(data,@kernel,nu)` and plot the resulting boundary with `meb_plot(data,meb)`. Which kernel seems to be the most appropriate for the given data? Rescale the data `data.X = data.X/4` and retry the kernels (and different values of ν). Why do some of the solutions change qualitatively in response to scaling?