

Problem Set 2 Solutions

Problem 1 Radial basis kernel

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

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

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

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

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 $\underline{\theta}^*$ as a function of the Lagrange multipliers. (this should not involve lengthy calculations).

The Lagrangian for this optimization problem is:

$$\begin{aligned} L(\underline{\theta}, \alpha) &= \frac{1}{2}\|\underline{\theta}\|^2 - \sum_{i=1}^n \alpha_i (y_i \underline{\theta} \cdot \underline{\phi}(\underline{x}_i) - 1) \\ &= \frac{1}{2}\|\underline{\theta}\|^2 - \underline{\theta} \cdot \left(\sum_{i=1}^n \alpha_i y_i \underline{\phi}(\underline{x}_i) \right) + \sum_{i=1}^n \alpha_i \end{aligned}$$

Here each α_i is *unconstrained*, because we have equality constraints rather than inequality constraints.

As usual, the dual optimization problem is $\max_{\alpha} g(\alpha) = \max_{\alpha} \min_{\underline{\theta}} L(\underline{\theta}, \alpha)$. For a fixed α (namely some optimal α^*), $L(\underline{\theta}, \alpha)$ is positively quadratic in $\underline{\theta}$. We can obtain the optimal $\underline{\theta}^*$ from the first-order condition $\frac{\partial L(\underline{\theta}, \alpha)}{\partial \underline{\theta}} = 0$:

$$\underline{\theta}^* = \sum_{j=1}^n \alpha_j^* y_j \phi(\underline{x}_j)$$

For convenience, we will use the short-hand $\underline{\theta}^* = \Phi(\underline{y} \bullet \underline{\alpha}^*)$. Here \bullet represents an element-wise product and Φ is a $d \times n$ matrix, where the i^{th} column is $\phi(\underline{x}_i)$. (Of course, $d = \infty$ for the RBF kernel.)

(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.

Our constraints are equivalent to:

$$\phi(\underline{x}_i)^T \underline{\theta} = y_i, \quad i = 1, \dots, n \tag{3}$$

Using matrix short-hand notation and substituting $\underline{\theta}^* = \Phi(\underline{y} \bullet \underline{\alpha}^*)$, we obtain:

$$\begin{aligned} \Phi^T \underline{\theta} &= \underline{y} \\ \Phi^T \Phi(\underline{y} \bullet \underline{\alpha}^*) &= \underline{y} \\ K(\underline{y} \bullet \underline{\alpha}^*) &= \underline{y} \end{aligned}$$

where $K = \Phi^T \Phi$ denotes the Gram matrix.

(c) Indicate briefly how we can use the following Michelli theorem to show that any n by n RBF 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 $\underline{x}_i, i = 1, \dots, n$.

Note that $\rho(t) = \exp\{-\frac{1}{2\sigma^2}t^2\}$ is a monotonic function in $t \in [0, \infty)$. Using the Michelli theorem, for any distinct set of points $\underline{x}_i, i = 1, \dots, n$, the matrix K , with entries $K_{ij} = \exp\{-\frac{1}{2\sigma^2}\|\underline{x}_i - \underline{x}_j\|^2\}$, is invertible.

(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.

As we have a distinct set of points, K is invertible. Then the linear system $K(\underline{y} \bullet \underline{\alpha}^*) = \underline{y}$ is feasible, and has a unique solution given by $\underline{\alpha}^* = \underline{y} \bullet (K^{-1}\underline{y})$. Therefore, $\underline{\theta}^* = \Phi(\underline{y} \bullet \underline{\alpha}^*) = \Phi K^{-1}\underline{y}$.

(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

$$\phi(\underline{x}) \cdot \phi(\underline{x}) = K(\underline{x}, \underline{x}) = 1 \tag{4}$$

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?

As $\sigma \rightarrow 0$, the points become very far apart with respect to σ , and our kernel matrix $K \rightarrow I$, the identity matrix. Because our constraints dictate that $K(\underline{y} \bullet \underline{\alpha}^*) = \underline{y}$, then $\underline{\alpha}^* \rightarrow \underline{1}$, the all-ones vector. Therefore,

$$\|\underline{\theta}^*\|^2 = \underline{\alpha}^{*T} K \underline{\alpha}^* \rightarrow \underline{1}^T I \underline{1} = n,$$

and we obtain a margin of $\frac{1}{\sqrt{n}}$ in the limit.

As $\sigma \rightarrow 0$, we can intuitively think of the kernel centered on \underline{x}_i , $K(\cdot, \underline{x}_i)$, as becoming a delta function $\delta(\cdot, \underline{x}_i)$. So consider the n -dimensional feature mapping $\phi(\cdot)$, where the i^{th} component feature is $\delta(\cdot, \underline{x}_i)$. In effect, the i^{th} data point then becomes the i^{th} standard basis vector \underline{e}_i in the limit. The geometric margin for the set of standard basis vectors $\{\underline{e}_i\}$ is $\frac{1}{\sqrt{n}}$, as we showed in HW1, Problem 1d.

(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”.

The simplest example to create is a set of 2 distinct points \underline{x} and \underline{x}' , both labeled +1. Denote $k = K(\underline{x}, \underline{x}') = \exp\{-\frac{1}{2\sigma^2}\|\underline{x} - \underline{x}'\|^2\}$. The Gram matrix can be written as:

$$K = \begin{bmatrix} 1 & k \\ k & 1 \end{bmatrix}$$

Solving the system $K(\underline{1} \bullet \alpha) = \underline{1}$ yields $\alpha = [\frac{1}{k+1} \quad \frac{1}{k+1}]^T$ and $\|\underline{\theta}^*\|^2 = \alpha K \alpha = \frac{2}{k+1}$. Therefore, the margin is $\sqrt{\frac{k+1}{2}}$. As long as \underline{x} and \underline{x}' are distinct, we have that $k > 0$, so the margin is always greater than $\frac{1}{\sqrt{2}}$.

As we take the 2 points arbitrarily close together (or alternatively, imagine that $\sigma \rightarrow \infty^+$), then $k \rightarrow 1$, and we obtain a margin of 1. Is 1 always the largest possible margin that we can obtain? For the RBF kernel, one can think of the corresponding infinite-dimensional feature vectors $\phi(\underline{x}_i)$ as lying on the unit ball, as they are unit-normalized: $\|\phi(\underline{x}_i)\|^2 = K(\underline{x}_i, \underline{x}_i) = 1$. So yes, the largest possible margin must be 1.

Intuitively, as $\sigma \rightarrow \infty^+$, kernels centered at distinct points gradually become indistinguishable. In effect, all the feature vectors collapse onto each other (to a single point) on the unit ball. As they do, the margin goes to 1 in the limit.

Problem 2 Anomaly detection

(a) For both the anomaly detection problem and the minimum enclosing ball problem $\underline{\theta}^* = \text{sum}_{i=1}^n \alpha_i^* \phi(\underline{x}_i)$. If we compare the dual optimization problems and we get the same α^* values when is $\|\phi(\underline{x}_i)\| = c$ for $i = 1, \dots, n$ then both problems are identical in that case.

For the anomaly detection problem, the dual is:

$$\max_{\alpha} - \sum_{i,j=1}^n \alpha_i \alpha_j K(\underline{x}_i, \underline{x}_j) \quad \text{subject to} \quad 0 \leq \alpha_i \leq \frac{1}{\nu n}, \quad i = 1, \dots, n, \quad \sum_{i=1}^n \alpha_i = 1 \quad (5)$$

For the MEB problem, the dual is:

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

We know that $K(\underline{x}_i, \underline{x}_i) = \|\phi(\underline{x}_i)\|^2 = c^2$ and $\sum_{i=1}^n \alpha_i = 1$, therefore $\sum_{i=1}^n \alpha_i K(\underline{x}_i, \underline{x}_i) = c^2$. This tells us that the term $\sum_{i=1}^n \alpha_i K(\underline{x}_i, \underline{x}_i)$ has no effect in optimization problem. Then we get the dual of the MEB in the case where $\|\phi(\underline{x}_i)\| = c$ is equivalent to the dual of the anomaly detection and both problem gives us the same α^* , hence the same $\underline{\theta}^*$.

(b) If we are given α^* , we know that any support vector, x_i , on the margin has to satisfy $R^2 = \|\theta - \underline{x}_i\|^2$. We find such a SV by taking any α_i such that $0 < \alpha_i < \frac{1}{\nu n}$ and get $R = \sqrt{\|\theta - \underline{x}_i\|^2}$.

If we are given only $\underline{\theta}(\alpha^*)$ we need to find:

$$R = \underset{R}{\operatorname{argmin}} R^2 + \frac{1}{\nu n} \sum_{i=1}^n \xi_i \quad \text{subject to} \quad \|\underline{\theta} - \underline{\phi}(\underline{x}_i)\|^2 \leq R^2 + \xi_i, \xi_i \geq 0 \text{ for all } i = 1, \dots, n \quad (7)$$

By expressing the loss-function in term of R we get:

$$R = \underset{R}{\operatorname{argmin}} R^2 + \frac{1}{\nu n} \sum_{i=1}^n \max(0, \|\underline{\theta} - \underline{\phi}(\underline{x}_i)\|^2 - R^2) \quad (8)$$

This is a simple convex problem with the only variable R that we can find a solution to easily.

(c) build_meb.m

```
function meb = meb_build(data, kernel, nu)

X = data.X;
n = size(X,1);

if (nargin<3), nu = 1/n; end; % no points omitted if nu left unspecified

% initialize the gram matrix
A = feval(kernel,X,X);
A = (A+A')/2; % symmetrize again for numerical reasons
alpha_max = 1/(nu*n); % cap on alpha's

% solve dual problem...
%options = optimset('MaxIter',1000000000);
alpha = quadprog(2*A,-diag(A), [], [], ones(1,n), 1, zeros(n,1), repmat(alpha_max,n,1));

% norm^2 of the primal theta parameters
meb.norm2theta = alpha'*A*alpha;

% select support vectors
S = find(1e-8 < alpha);

% calculate radius^2
index=find(alpha(S)<alpha_max); Adia = diag(A);
meb.R2 = max( meb.norm2theta - 2*A(S(index),S)*alpha(S) + Adia(S(index)) );

meb.kernel = kernel;
meb.alpha = alpha(S); % keep only support vectors
meb.XS = X(S,:); % store support vectors
meb.nu = nu;
```

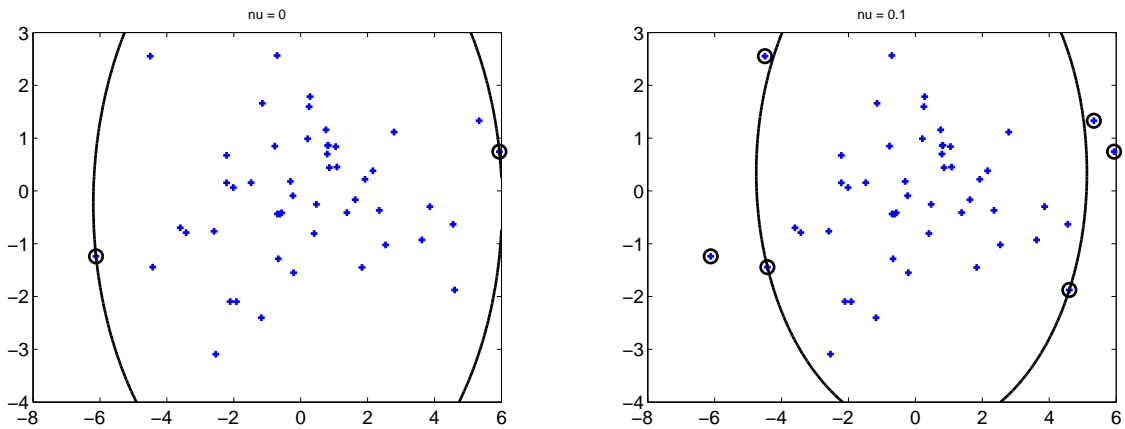
meb_discriminant_function.m

```
function f = meb_discrim_func(X,meb)

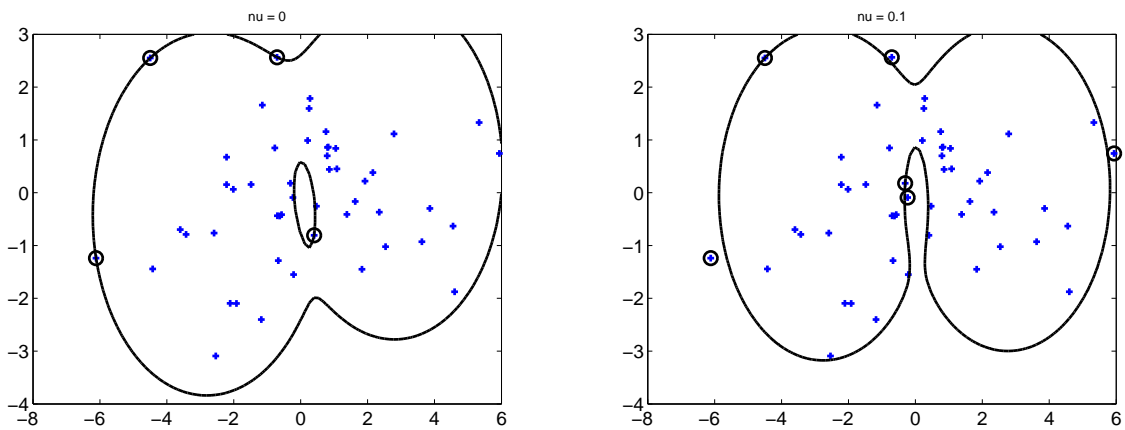
f = meb.norm2theta - 2*feval(meb.kernel,X,meb.XS)*meb.alpha;
for i=1:size(X,1),
    f(i) = f(i) + feval(meb.kernel,X(i,:),X(i,:));
end;
f = meb.R2-f; % R^2 - |theta - X|^2
```

(d) K2 seems most appropriate given the data, it better approximate the underlying distribution better. K1 is too simple (underfit) and Kr is too complex (overfit).

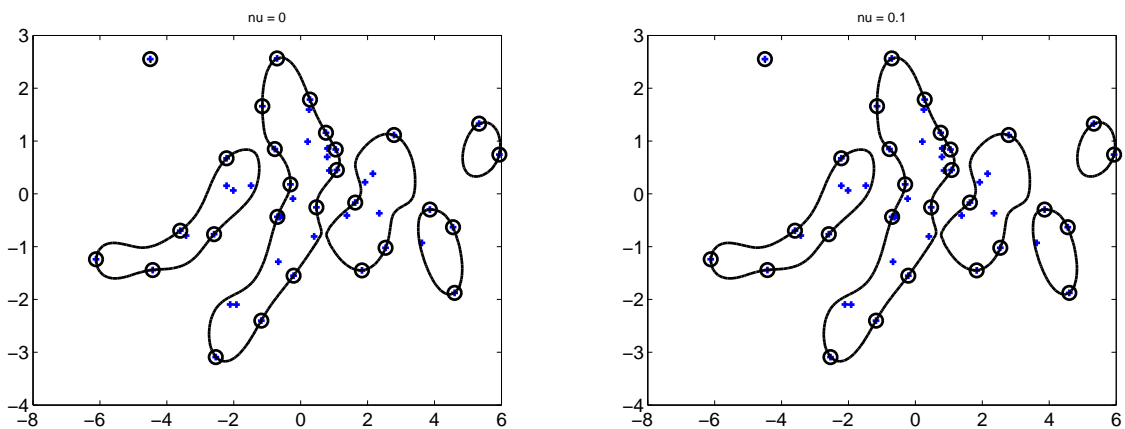
After rescaling the data by dividing by four, Kr is affected the most. It gives a smoother solution since rescaling the data is equivalent to increasing σ , making Kr to decay slower.



(a) Linear Kernel K1

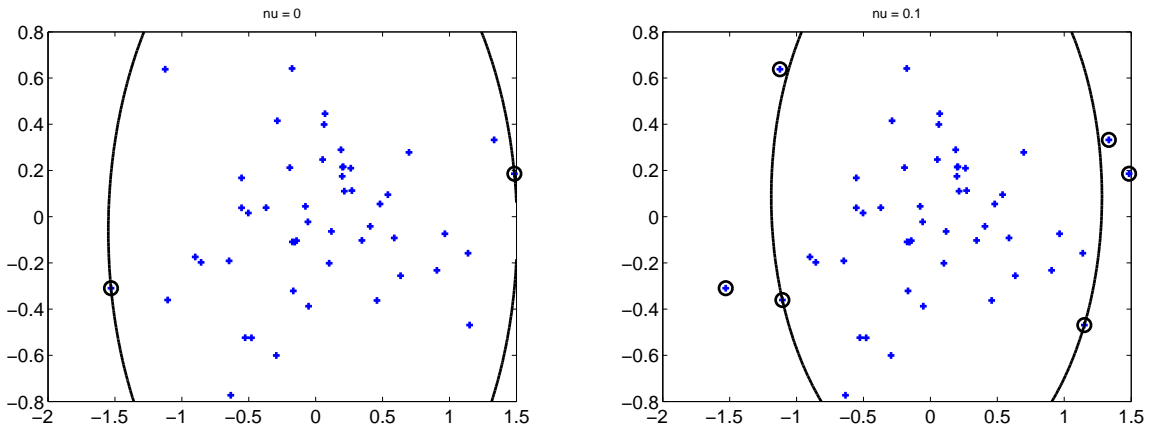


(b) Quadratic kernel K2

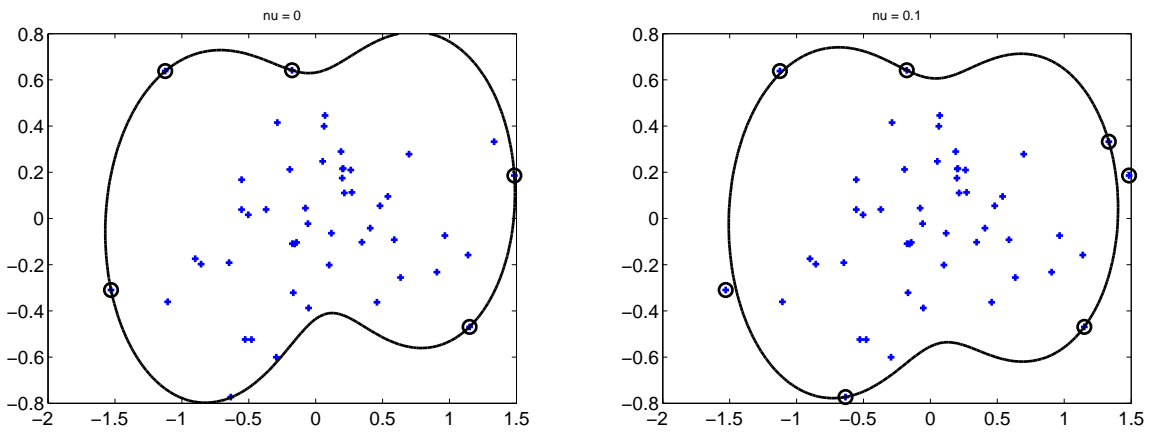


(c) Radial basis kernel K_r

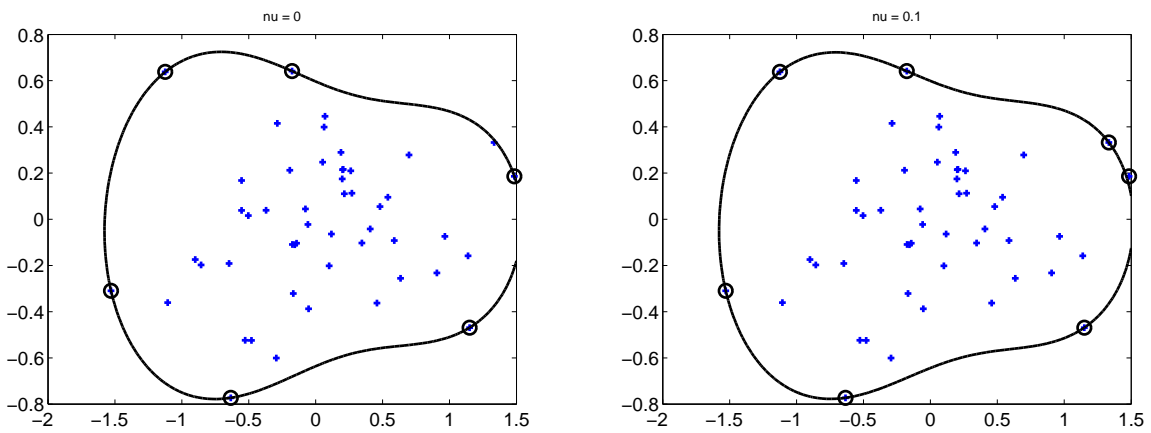
Figure 1: Decision boundary of data



(a) Linear Kernel K1



(b) Quadratic kernel K2



(c) Radial basis kernel Kr

Figure 2: Decision boundary of data/4