Mehryar Mohri
Foundations of Machine Learning 2015
Courant Institute of Mathematical Sciences
Homework assignment 2
October 23, 2015
Due: November 09, 2015

A. VC-dimension of convex combinations

1. Let H be a family of functions mapping from an input space $\mathcal{X}$ to $\{-1, +1\}$ and let T be a positive integer. Give an upper bound on the VC-dimension of the family of functions $\mathcal{F}_T$ defined by

$$\mathcal{F} = \left\{ \operatorname{sgn}\left(\sum_{t=1}^{T} \alpha_t h_t \right) : h_t \in H, \alpha_t \geq 0, \sum_{t=1}^{T} \alpha_t \leq 1 \right\}.$$

(*Hint*: you can use Problem C. of (Foundations of Machine Learning, HW2, 2014, `http://www.cs.nyu.edu/~mohri/ml14/hw2.pdf` and its solution).

Solution: Following the hint, we can think of this family of functions as a one hidden layer neural network, where the hidden layer is represented by the functions $h_t \in H$, and the top layer is a threshold function characterized by $(\alpha_1, \ldots, \alpha_T)$. Denote this class of threshold functions by Δ_T. By problem C1 from FML2014-HW2, we can bound the growth function of $\mathcal{F}_T$ by:

$$\Pi_{\mathcal{F}_T}(m) \leq \Pi_{\Delta_T}(m) \left(\Pi_H(m)\right)^T.$$

By problem C3 from FML2014-HW2, the VC dimension of Δ_T is at most T, and we may further denote the VC dimension of H by d. Applying Sauer's lemma to the growth function yields:

$$\Pi_{\Delta_T}(m) \leq \left(\frac{em}{T}\right)^T, \quad \Pi_H(m) \leq \left(\frac{em}{d}\right)^d.$$

Thus, we have that

$$\Pi_{\mathcal{F}_T}(m) \leq \left(\frac{em}{T}\right)^T \left(\frac{em}{d}\right)^{Td}.$$

Finally, we may apply the hint in problem C2 from FML2014-HW2 with $m = \max\{4T \log_2(2e), 2Td \log_2(eT)\}$ to see that

$$\left(\frac{em}{T}\right)^T \left(\frac{em}{d}\right)^{Td} < 2^{4T \log_2(2e) + 2Td \log_2(eT)},$$

so that the VC Dimension of $\mathcal{F}_T$ is bounded by:

$$2T(2\log_2(2e) + d\log_2(eT)).$$

Note that a coarser but relatively simpler bound would be to write:

$$\left(\frac{em}{T}\right)^T \left(\frac{em}{d}\right)^{Td} < (em)^{T(d+1)},$$

and to apply the hint in problem C2 from FML2014-HW2 with $m = 2T(d+1)\log_2(eT(d+1))$. Notice that this is actually asymptotically optimal in T and d up to log terms.

B. Growth function

1. A *linearly separable labeling* of a set X of vectors in $\mathbb{R}^d$ is a classification of X into two sets X^+ and X^- with $X^+ = \{\mathbf{x} \in X : \mathbf{w} \cdot \mathbf{x} > 0\}$ and $X^- = \{\mathbf{x} \in X : \mathbf{w} \cdot \mathbf{x} < 0\}$ for some $\mathbf{w} \in \mathbb{R}^d$.

   Let $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ be a subset of $\mathbb{R}^d$.

   (a) Let $\{X^+, X^-\}$ be a dichotomy of X and let $\mathbf{x}_{m+1} \in \mathbb{R}^d$. Show that $\{X^+ \cup \{\mathbf{x}_{m+1}\}, X^-\}$ and $\{X^+, X^- \cup \{\mathbf{x}_{m+1}\}\}$ are linearly separable by a hyperplane going through the origin if and only if $\{X^+, X^-\}$ is linearly separable by a hyperplane going through the origin and $\mathbf{x}_{m+1}$.

   Solution: $\{X^+ \cup \{\mathbf{x}_{m+1}\}, X^-\}$ and $\{X^+, X^- \cup \{\mathbf{x}_{m+1}\}\}$ are linearly separable by a hyperplane going through the origin if and only if there exists $\mathbf{w}_1 \in \mathbb{R}^d$ such that

$$\forall \mathbf{x} \in X^+, \mathbf{w}_1 \cdot \mathbf{x} > 0 \quad \forall \mathbf{x} \in X^-, \mathbf{w}_1 \cdot \mathbf{x} < 0, \text{ and } \mathbf{w}_1 \cdot \mathbf{x}_{m+1} > 0 \tag{1}$$

   and there exists $\mathbf{w}_2 \in \mathbb{R}^d$ such that

$$\forall \mathbf{x} \in X^+, \mathbf{w}_2 \cdot \mathbf{x} > 0 \quad \forall \mathbf{x} \in X^-, \mathbf{w}_2 \cdot \mathbf{x} < 0, \text{ and } \mathbf{w}_2 \cdot \mathbf{x}_{m+1} < 0. \tag{2}$$

   For any $\mathbf{w}_1, \mathbf{w}_2$, the function $f\colon (t \mapsto t\mathbf{w}_1 + (1-t)\mathbf{w}_2) \cdot \mathbf{x}_{m+1}$ is continuous over $[0, 1]$. (1) and (2) hold iff $f(0) < 0$ and $f(1) > 0$, that is iff there exists $\mathbf{w} = t_0\mathbf{w}_1 + (1-t_0)\mathbf{w}_2$ linearly separating $\{X^+, X^-\}$ and such at $\mathbf{w} \cdot \mathbf{x}_{m+1} = 0$. □

(b) Let $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ be a subset of $\mathbb{R}^d$ such that any k-element subset of X with $k \leq d$ is linearly independent. Then, the number of linearly separable labelings of X is $C(m,d) = 2\sum_{k=0}^{d-1}\binom{m-1}{k}$. (*Hint*: prove by induction that $C(m+1,d) = C(m,d) + C(m,d-1)$.

Solution:

Repeating the formula, we obtain $C(m,d) = \sum_{k=0}^{m-1}\binom{m-1}{k}C(1,d-k)$. Since, $C(1,n) = 2$ if $n \geq 1$ and $C(1,n) = 0$ otherwise, the result follows. □

(c) Let $f_1, \ldots, f_p$ be p functions mapping $\mathbb{R}^d$ to $\mathbb{R}$. Define $\mathcal{F}$ as the family of classifiers based on linear combinations of these functions:

$$\mathcal{F} = \left\{ x \mapsto \operatorname{sgn}\left(\sum_{k=1}^{p} a_k f_k(x)\right) : a_1, \ldots, a_p \in \mathbb{R} \right\}.$$

Define Ψ by $\Psi(x) = (f_1(x), \ldots, f_p(x))$. Assume that there exists $x_1, \ldots, x_m \in \mathbb{R}^d$ such that every p-subset of $\{\Psi(x_1), \ldots, \Psi(x_m)\}$ is linearly independent. Then, show that

$$\Pi_{\mathcal{F}}(m) = 2\sum_{i=0}^{p-1}\binom{m-1}{i}.$$

Solution:

This is a direct application of the result of the previous question. □

C. Support Vector Machines

1. Download and install the `libsvm` software library from:

   `http://www.csie.ntu.edu.tw/~cjlin/libsvm/` ,

   and briefly consult the documentation to become more familiar with the tools.

2. Consider the `splice` data set

   `http://www.cs.toronto.edu/~delve/data/splice/desc.html`.

Download the already formatted training and test files of a noisy version of that dataset from

`http://www.cs.nyu.edu/~mohri/ml15/splice_noise_train.txt`
`http://www.cs.nyu.edu/~mohri/ml15/splice_noise_test.txt`.

Use the `libsvm` scaling tool to scale the features of all the data. The scaling parameters should be computed only on the training data and then applied to the test data.

3. Consider the corresponding binary classification which consists of distinguishing two types of splice junctions in DNA sequences using about 60 features. Use SVMs combined with polynomial kernels to tackle this problem.

   To do that, randomly split the training data into ten equal-sized disjoint sets. For each value of the polynomial degree, $d = 1, 3, 5$, plot the average cross-validation error plus or minus one standard deviation as a function of C (let other parameters of polynomial kernels in `libsvm` be equal to their default values), varying C in powers of 5, starting from a small value $C = 5^{-k}$ to $C = 5^k$, for some value of k. k should be chosen so that you see a significant variation in training error, starting from a very high training error to a low training error. Expect longer training times with `libsvm` as the value of C increases.

   Solution:

   Figure 1 shows the average cross-validation performance as a function of the regularization parameter C. Note that the algorithm starts to exhibit some over-fitting as C becomes very large. The performance for several choices of d and C are essentially indistinguishable; one suitable choice of optimal parameters is $C^* = 5^1$ and $d^* = 3$. □

4. Let (C^*, d^*) be the best pair found previously. Fix C to be C^*. Plot the ten-fold cross-validation error and the test errors for the hypotheses obtained as a function of d. Plot the average number of support vectors obtained as a function of d. How many of the support vectors lie on the marginal hyperplanes? Plot the soft margin of the solution as a function of d.

   Solution:

   The first plot in Figure 4 compares CV and test errors for C^* as a function of d. As expected test error is slightly higher than CV error.

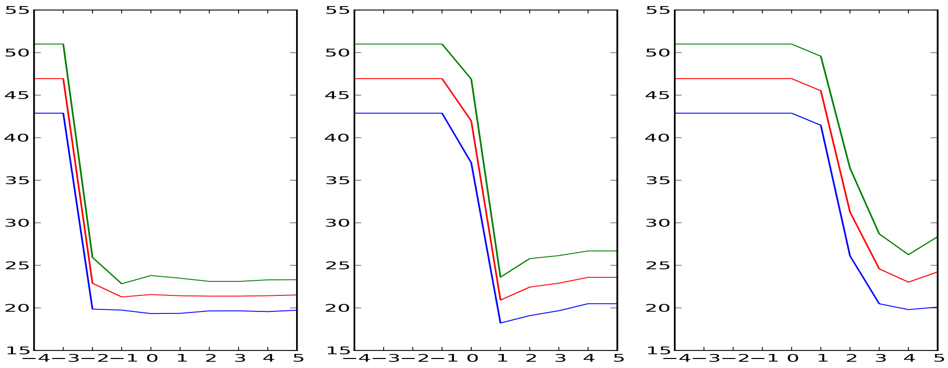

Figure 1: Average error (red) according to 10-fold cross-validation, with error-bars (green and blue) indicating one standard deviation. Left, middle and right panels correspond to $d = 1, 3, 5$ respectively. On the x-axis we have $\log_5 C$.

The second plot shows that the total number of support vectors and the number of support vectors on the marginal hyperplanes. The last plot shows the margin as a function of d. □

5. Now, combine SVMs with Gaussian kernels to tackle the same task. Use cross-validation as before to determine the best value of C and σ, varying C in powers of 5, and σ in powers of 2 for a reasonable range so that you see a significant variation in training error, as before. Fix C and σ to the best values found via cross-validation. How does the test error of the solution compare to the best result obtained using polynomial kernels? What is the value of the soft margin?

   Solution:

   Figure 3 shows the average cross-validation performance as a function of the regularization parameter C and $g = 1/\sigma^2$. $C^* = 1$ and $g^* = 0.03125$. Figure 4 shows the test and validation error as a function of degree (top panel), the number of total and marginal support vectors (second panel) and the margin as function of g.

   Note that the best test error for polynomial kernels is around 23% and with Gaussian kernels it is around 19%. □

6. Here, use as a kernel the sum of the best polynomial kernel (degree

d^*) and the Gaussian kernel with the best parameter σ you found in the previous question. Use cross-validation as before to determine the best value of C. How does the test error of the solution compare to the best result obtained in the previous questions?

Solution: There are multiply ways to solve this problem. If you are using `libsvm` in `MATLAB`, then you can simply precompute the kernel matrix and use it to solve the problem. Otherwise, you can directly modify `libsvm` code (see `libsvm` FAQ on how to do this). In our case, we use Gaussian kernel with $g = 0.03125$ and polynomial kernel with $d = 3$. The test error is around 21% which is slightly worse than for Gaussian kernels, but better than for polynomial kernels.

D. Kernels

Show that the following kernels are PDS.

1. Let n be a positive integer. K is defined by $K(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{N} \cos^n(x_i^2 - y_i^2)$ for all $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^N \times \mathbb{R}^N$.

   Solution: Since the product and sum of PDS kernels is PDS, it suffices to show that $k \colon x \mapsto \cos(x^2 - y^2)$ is PDS over $\mathbb{R} \times \mathbb{R}$. This is clear since

   $$k(x, y) = \cos(x^2)\cos(y^2) + \sin(x^2)\sin(y^2) = \Phi(x) \cdot \Phi(y),$$

   with

   $$\Phi(x) = \begin{bmatrix} \cos(x^2) \\ \sin(x^2) \end{bmatrix}.$$

   □

2. Let σ be a positive real number. K is defined by $K(x, y) = e^{-\frac{\|\mathbf{x}-\mathbf{y}\|}{\sigma}}$ for all $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^N \times \mathbb{R}^N$ (*Hint*: you could show that K is the normalized kernel of a kernel K' and show that K' is PDS using the following equality: $\|\mathbf{x} - \mathbf{y}\| = \frac{1}{2\Gamma(\frac{1}{2})} \int_0^{+\infty} \frac{1 - e^{-t\|\mathbf{x}-\mathbf{y}\|^2}}{t^{\frac{3}{2}}} \, dt$ valid for all $\mathbf{x}, \mathbf{y}$).

   Solution: It suffices to show that K is the normalized kernel associated to the kernel K' defined by

   $$\forall (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^N \times \mathbb{R}^N, K'(\mathbf{x}, \mathbf{y}) = e^{\phi(\mathbf{x}, \mathbf{y})}$$

where $\phi(\mathbf{x},\mathbf{y}) = \frac{1}{\sigma}[\|\mathbf{x}\| + \|\mathbf{y}\| - \|\mathbf{x}-\mathbf{y}\|]$, and to show that K' is PDS. For the first part, observe that

$$\frac{K'(\mathbf{x},\mathbf{y})}{\sqrt{K'(\mathbf{x},\mathbf{x})K'(\mathbf{y},\mathbf{y})}} = e^{\phi(\mathbf{x},\mathbf{y})-\frac{1}{2}\phi(\mathbf{x},\mathbf{x})-\frac{1}{2}\phi(\mathbf{y},\mathbf{y})} = e^{-\frac{\|\mathbf{x}-\mathbf{y}\|}{\sigma}}.$$

To show that K' is PDS, it suffices to show that ϕ is PDS, since composition with a power series with non-negative coefficients (here exp) preserve the PDS property. Now, for any $c_1, \ldots, c_n \in \mathbb{R}$, let $c_0 = -\sum_{i=1}^n c_i$, then, we can write

$$\begin{aligned}\sum_{i,j=1}^n c_i c_j \phi(\mathbf{x}_i,\mathbf{x}_j) &= \frac{1}{\sigma}\sum_{i,j=1}^n c_i c_j [\|\mathbf{x}_i\| + \|\mathbf{x}_j\| - \|\mathbf{x}_i - \mathbf{x}_j\|] \\ &= \frac{1}{\sigma}\Big[-\sum_{i=1}^n c_0 c_i \|\mathbf{x}_i\| + -\sum_{i=1}^n c_0 c_j \|\mathbf{x}_j\| - \sum_{i,j=1}^n c_i c_j \|\mathbf{x}_i - \mathbf{x}_j\|\Big] \\ &= -\frac{1}{\sigma}\sum_{i,j=0}^n c_i c_j \|\mathbf{x}_i - \mathbf{x}_j\|,\end{aligned}$$

with $\mathbf{x}_0 = 0$. Now, for any $z \in \mathbb{R}$, the following equality holds:

$$z^{\frac{1}{2}} = \frac{1}{2\Gamma(\frac{1}{2})}\int_0^{+\infty} \frac{1-e^{-tz}}{t^{\frac{3}{2}}}\, dt.$$

Thus,

$$\begin{aligned}-\frac{1}{\sigma}\sum_{i,j=0}^n c_i c_j \|\mathbf{x}_i - \mathbf{x}_j\| &= \frac{1}{2\Gamma(\frac{1}{2})}\int_0^{+\infty} -\frac{1}{\sigma}\sum_{i,j=0}^n c_i c_j \frac{1-e^{-t\|\mathbf{x}_i-\mathbf{x}_j\|^2}}{t^{\frac{3}{2}}}\, dt \\ &= \frac{1}{2\Gamma(\frac{1}{2})}\int_0^{+\infty} \frac{1}{\sigma}\frac{\sum_{i,j=0}^n c_i c_j e^{-t\|\mathbf{x}_i-\mathbf{x}_j\|^2}}{t^{\frac{3}{2}}}\, dt.\end{aligned}$$

Since a Gaussian kernel is PDS, the inequality $\sum_{i,j=0}^n c_i c_j e^{-t\|\mathbf{x}_i-\mathbf{x}_j\|^2} \geq 0$ holds and the right-hand side is non-negative. Thus, the inequality $-\frac{1}{\sigma}\sum_{i,j=0}^n c_i c_j \|\mathbf{x}_i - \mathbf{x}_j\| \geq 0$ holds, which shows that ϕ is PDS. □

Alternatively, one can also apply the theorem on page 43 of the lecture slides on kernel methods to reduce the problem to showing that the norm $G(x,y) = \|x-y\|$ is a NDS function. This can be shown through a direct application of the definition of NDS together with the representation of the norm given in the hint.

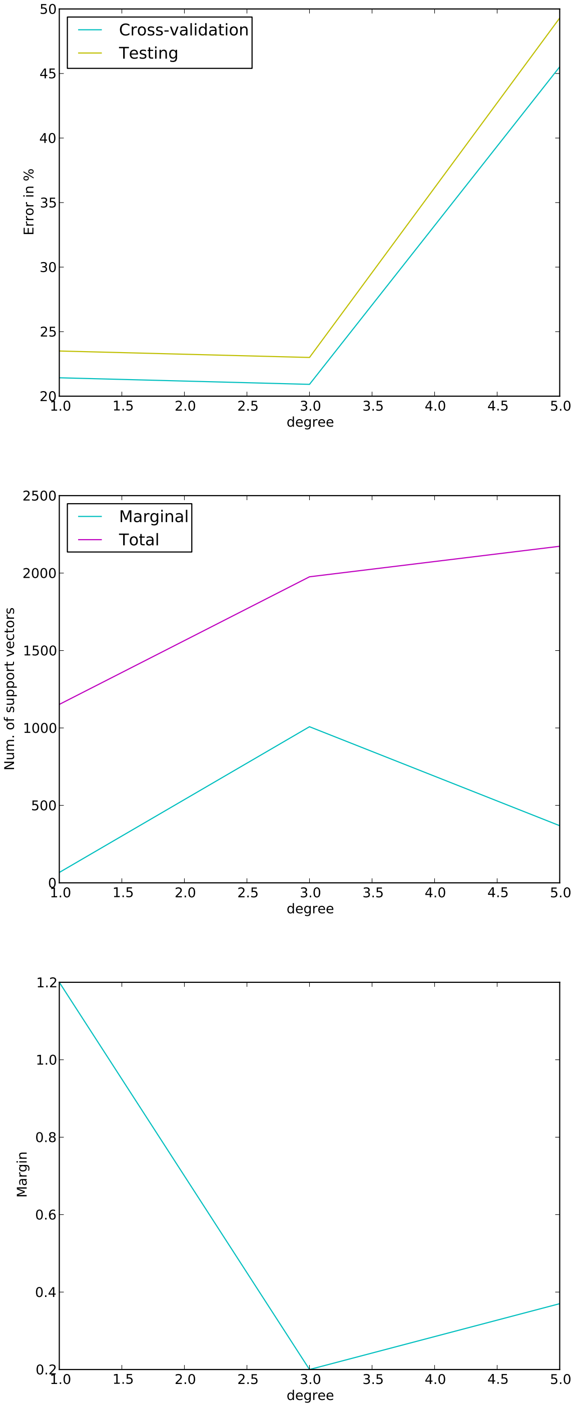

Figure 2: The test and validation error as a function of d (top panel), the number of total and marginal support vectors (second panel) and the margin as function of d.

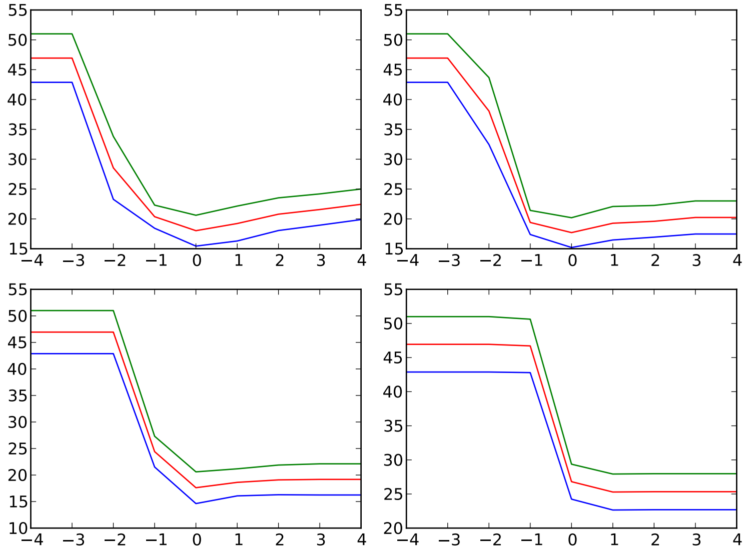

Figure 3: Average error (red) according to 10-fold cross-validation, with error-bars (green and blue) indicating one standard deviation. Top left, top right, bottom left and bottom right correspond to $g = 0.015625, 0.03125, 0.0625, 0.125$ respectively. Note that $g = 1/2\sigma^2$. On the x-axis we have $\log_5 C$.

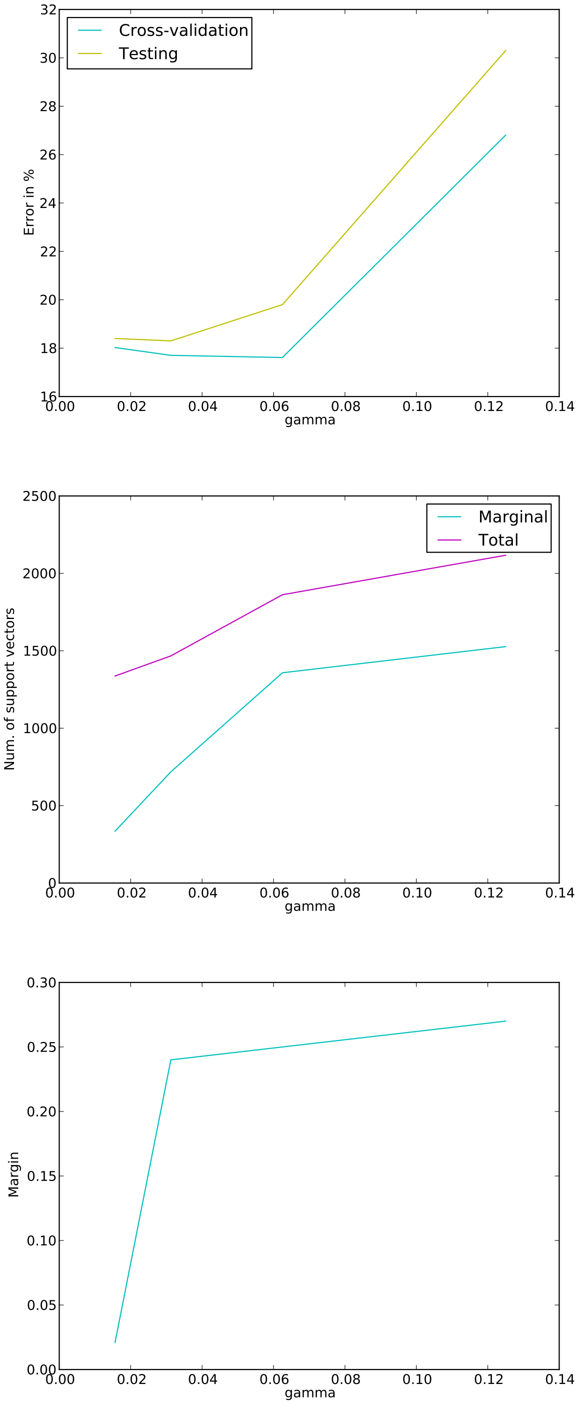

Figure 4: The test and validation error as a function of degree (top panel), the number of total and marginal support vectors (second panel) and the margin as function of g.