582669 Supervised Machine Learning (Spring 2011)

Course examination, solutions (Jyrki Kivinen)

General comment: The exam turned out to be more difficult than intended (in particular, too long). As partial compensation, the exam points were multiplied by 1.2 in the grading of the course.

1. This is directly from Homework 1(a).

2. (a) This is from pages 87–88 of the lecture notes. A complete answer should also include a definition of "margin."

   (b) This is given on pages 91–93 of the lecture notes.

3. (a) We formulate the problem as follows:

   Variables: $\boldsymbol{w} \in \mathbb{R}^d$, $r \in \mathbb{R}$

   minimise R

   subject to $\|\boldsymbol{w} - \boldsymbol{x}_i\|_2^2 - R \leq 0$ for $i = 1, \dots, m$.

   Notice that we have used the squared radius $R = r^2$ to make the problem convex.

   To obtain the dual, we write the Lagrangian

$$L(\boldsymbol{w}, R, \boldsymbol{\alpha}) = R + \sum_{i=1}^{m} \alpha_i (\|\boldsymbol{w} - \boldsymbol{x}_i\|_2^2 - R)$$

   where $\alpha_i \geq 0$. To minimise with respect to the original variables, we calculate the derivatives

$$\begin{aligned} \frac{\partial L(\boldsymbol{w}, R, \boldsymbol{\alpha})}{\partial \boldsymbol{w}} &= 2 \sum_{i=1}^{m} \alpha_i (\boldsymbol{w} - \boldsymbol{x}_i) \\ \frac{\partial L(\boldsymbol{w}, R, \boldsymbol{\alpha})}{\partial R} &= 1 - \sum_{i=1}^{m} \alpha_i. \end{aligned}$$

   and set them to zero, getting

$$\begin{aligned} \boldsymbol{w} &= \sum_{i=1}^{m} \alpha_i \boldsymbol{x}_i \\ \sum_{i=1}^{m} \alpha_i &= 1. \end{aligned}$$

   (Notice that together with the constraints $\alpha_i \geq 0$ these equations imply that the centre $\boldsymbol{w}$ is inside the convex hull of the points $\boldsymbol{x}_i$, which seems intuitive.) Substituting this into the Lagrangian we get

$$\begin{aligned}
L(\boldsymbol{w}, R, \boldsymbol{\alpha}) &= R + \sum_{i=1}^{m} \alpha_i (\|\boldsymbol{w} - \boldsymbol{x}_i\|_2^2 - R) \\
&= \sum_{i=1}^{m} \alpha_i (\boldsymbol{w} \cdot \boldsymbol{w} - 2\boldsymbol{w} \cdot \boldsymbol{x}_i + \boldsymbol{x}_i \cdot \boldsymbol{x}_i) \\
&= \boldsymbol{w} \cdot \boldsymbol{w} - 2\boldsymbol{w} \cdot \boldsymbol{w} + \sum_{i=1}^{m} \alpha_i \boldsymbol{x}_i \cdot \boldsymbol{x}_i \\
&= \sum_{i=1}^{m} \alpha_i \boldsymbol{x}_i \cdot \boldsymbol{x}_i - \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j \boldsymbol{x}_i \cdot \boldsymbol{x}_j.
\end{aligned}$$

Hence, the dual function is

$$G(\boldsymbol{\alpha}) = \sum_{i=1}^{m} \alpha_i \boldsymbol{x}_i \cdot \boldsymbol{x}_i - \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j \boldsymbol{x}_i \cdot \boldsymbol{x}_j,$$

and the dual problem is maximising this under the constraints $\alpha_i \geq 0$ and $\sum_i \alpha_i = 1$. (Notice that by complementary slackness, we have $\alpha_i \neq 0$ only when $\|\boldsymbol{w} - \boldsymbol{x_i}\|$ is exactly $\sqrt{R}$. Hence, moving points $\boldsymbol{x}_i$ inside the interior of the ball does not change the solution, which again is intuitively correct.) Suppose now that the instances are actually feature vectors, so $\boldsymbol{x}_i = \boldsymbol{\psi}(z_i)$ for some z_i. Here $\boldsymbol{\psi}$ is a feature map, for which we assume the corresponding kernel function is k. The dual function now becomes

$$G(\boldsymbol{\alpha}) = \sum_{i=1}^{m} \alpha_i k(x_i, x_i) - \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j k(x_i, x_j)$$

and the constraints remain the same. Thus, we can solve the dual without explicitly computing any feature vectors. The solution in feature space is then

$$\boldsymbol{w} = \sum_{i=1}^{m} \alpha_i \boldsymbol{\psi}(z_i).$$

(b) For the soft version, we introduce for each constraint a slack variable ξ_i. Analogously to soft-margin SVM, the optimisation problem becomes

Variables: $\boldsymbol{w} \in \mathbb{R}^d$, $R \in \mathbb{R}$, $\xi_1, \ldots, \xi_m$

minimise $R + C \sum_{i=1}^{m} \xi_i$

subject to $\|\boldsymbol{w} - \boldsymbol{x}_i\|_2^2 - R - \xi_i \leq 0$ for $i = 1, \ldots, m$

$\xi_i \geq 0$ for $i = 1, \ldots, m$

where $C > 0$ is a parameter we choose in practice by cross-validation or some similar method.

4. (a) Now H is the class of monotone conjunctions over n variables.

Claim 1: $\mathrm{VCdim}(H) \leq n$.

Proof: There are exactly 2^n monotone conjunctions, since for each of the n variables we can choose to include it or not include it in the formula. (As noted in the problem, not including any variables gives the function that is identically $+1$.) Since always $\mathrm{VCdim}(H) \leq \log_2 |H|$, the claim follows. □

Claim 2: $\mathrm{VCdim}(H) \geq n$.

Proof: We construct a set of n elements $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ that is shattered by H. Let $z_{ii} = -1$ for all i, and $z_{ij} = 1$ when $i \neq j$. Consider any set $I \subseteq \{1, \ldots, n\}$. We need to show that there is a monotone conjunction f such that $f(\boldsymbol{z}_i) = 1$ if $i \in I$, and $f(\boldsymbol{z}_i) = 1$ if $i \notin I$. We choose $f = \wedge_{i \notin I} v_i$.

If $i \notin I$, then $f(\boldsymbol{z}) = -1$ for any instance $\boldsymbol{z}$ with $z_i = -1$. In particular, $f(\boldsymbol{z_i}) = -1$.

If $i \in I$, then v_i does not appear in the conjuntion f. Since for z_i we have $z_{ij} = 1$ for all $j \neq i$, we have in particular $z_{ij} = 1$ for all j such that v_j is included in the conjunction. Hence, $f(\boldsymbol{z_i}) = 1$. □

(b) There is a universal constant C such that the following holds: Assume that $\mathrm{VCdim}(H) = d < \infty$, and that there is some probability distribution P over $X \times Y$. Let $0 < \varepsilon, \delta \leq 1$. Assume we draw a sample of m points $((x_1, y_1), \ldots, (x_m, y_m))$ independently from P, where

$$m \geq \frac{C}{\varepsilon^2} \left(d \ln \frac{2}{\varepsilon} + \ln \frac{2}{\delta} \right).$$

Then with probability at least $1 - \delta$ we have

$$\left| R(h) - \hat{R}(h) \right| \leq \varepsilon$$

for all $h \in H$. Here R and $\hat{R}$ are the true and empirical risks for the discrete loss:

$$\begin{aligned} R(h) &= \mathrm{E}_{(x,y)\sim P}[L_{0-1}(y, h(x))] \\ \hat{R}(h) &= \frac{1}{m} \sum_{i=1}^{m} L_{0-1}(y_i, h(x_i))]. \end{aligned}$$