THE AUSTRALIAN NATIONAL UNIVERSITY
First Semester Examinations 2001 — Solutions Sheet

ENGN4520 ENGINEERING S1
INTRODUCTION TO MACHINE LEARNING

Problem 1 (Cancer Diagnosis, 10 Points)
Assume a patient visits a physician to undergo a test whether he has a certain type of cancer or not. The probability that an arbitrary person will develop this type of cancer is 1 in 100,000.

1. *The patient tests positive (= cancer) to a test with the following properties: in 99 out of 100 cases it detects cancer if it is present. However, it raises a false alarm in 2 out of 1000 cases.*

   What is the probability of cancer, given that the test is positive.

2. *Additionally, the patient exhibits a certain symptom that is typically associated with cancer. In particular, all cancer patients show this symptom, however in one of 500 cases another disease is responsible for this symptom. Assume that the presence of this symptom is* independent *of the outcome of the previous test.*

   What is the probability of cancer now, given the knowledge about the symptom and *the test.*

Solution

1. We apply Bayes' rule and find

$$\begin{aligned} & \Pr(\text{ cancer} \mid \text{test positive}) \\ = & \frac{\Pr(\text{ test positive} \mid \text{cancer}) \Pr(\text{ cancer})}{\Pr(\text{ test positive})} \\ = & \frac{\Pr(\text{ test positive} \mid \text{cancer}) \Pr(\text{ cancer})}{\Pr(\text{ test positive} \mid \text{cancer}) \Pr(\text{ cancer}) + \Pr(\text{ test positive} \mid \text{no cancer}) \Pr(\text{ no cancer})} \\ = & \frac{0.99 \cdot 10^{-5}}{0.99 \cdot 10^{-5} + 0.002 \cdot (1 - 10^{-5})} = 0.0049. \end{aligned}$$

   This means that the patient is not very likely to have cancer, given only the test.

2. The symptom can be treated just like another test, only with the difference that now the prior probability of someone having cancer is no more 10^{-5} but 0.0049. This yields (we ignore the additional conditioning in the notation)

$$\begin{aligned} & \Pr(\text{ cancer} \mid \text{symptom}) \\ = & \frac{\Pr(\text{ symptom} \mid \text{cancer}) \Pr(\text{ cancer})}{\Pr(\text{ symptom} \mid \text{cancer}) \Pr(\text{ cancer}) + \Pr(\text{ symptom} \mid \text{no cancer}) \Pr(\text{ no cancer})} \\ = & \frac{1 \cdot 0.0049}{1 \cdot 0.0049 + 0.002 \cdot (1 - 0.0049)} = 0.7112. \end{aligned}$$

   This means that most likely the patient has cancer. Note that it is not enough, to consider just the presence of the symptom without any other information.

Problem 2 (Weighting Patterns with Support Vectors, 15 Points)
Assume we have a linear Support Vector Machine with soft margin loss, i.e.

$$c(\mathbf{x}, y, f(\mathbf{x})) = \max(0, 1 - yf(\mathbf{x})) \tag{1}$$

which is to be trained on some training data $(\mathbf{x}_1, y_1), \ldots (\mathbf{x}_m, y_m)$*. Moreover assume that we know that some of the observations* $(\mathbf{x}_i, y_i)$ *are more important than others, specifically that there exist weighting coefficients* $C_i > 0$ *such that we minimize a modified regularized risk functional*

$$\sum_{i=1}^{m} C_i c(\mathbf{x}_i, y_i, f(\mathbf{x}_i)) + \frac{1}{2}\|w\|^2. \tag{2}$$

1. *Rewrite eq. (2) such that it becomes a constrained quadratic optimization problem, i.e. with linear constraints and a quadratic objective function.*
2. *Derive the Lagrange function corresponding to the constrained optimization problem.*
3. *Compute the dual optimization problem.*
4. *Compare the result to the standard soft margin Support Vector Machine.*

Solution

1. We know (see lecture notes) that we can rewrite $c(\mathbf{x}, y, f(\mathbf{x}))$ as the solution of the optimization problem

$$\begin{array}{ll} \underset{\xi}{\text{minimize}} & \xi \\ \text{subject to} & yf(\mathbf{x}) \geq 1 - \xi \text{ and } \xi \geq 0 \end{array}$$

   Therefore we obtain for the optimization problem over $C_i\xi_i$ and $\|\mathbf{w}\|^2$ (with $f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x}_i \rangle + b$)

$$\begin{array}{ll} \underset{\xi_i, \mathbf{w}, b}{\text{minimize}} & \displaystyle\sum_{i=1}^{m} C_i\xi_i + \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{subject to} & y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 - \xi_i \text{ and } \xi_i \geq 0 \text{ for all } 1 \leq i \leq m \end{array}$$

2. The Lagrange function is objective function plus constraints times Lagrange multipliers. This yields

$$L = \sum_{i=1}^{m} C_i\xi_i + \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{m} \alpha_i \left(1 - \xi_i - y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b)\right) + \sum_{i=1}^{m} \eta_i(-\xi_i) \text{ where } \alpha_i, \eta_i \geq 0$$

3. Partial derivatives with respect to $\mathbf{w}, \xi_i, b$ have to vanish due to the saddlepoint condition.

$$\begin{aligned} \partial_{\mathbf{w}} L &= \mathbf{w} - \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i \text{ and therefore } \mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i \\ \partial_b L &= -\sum_{i=1}^{m} \alpha_i y_i \text{ and therefore } \sum_{i=1}^{m} \alpha_i y_i = 0 \\ \partial_{\xi_i} L &= C_i - \alpha_i - \eta_i \text{ and therefore } \alpha_i \in [0, C_i] \text{ and } \eta_i = C_i - \alpha_i \end{aligned}$$

   Substituting the saddlepoint equations into L yields the dual optimization problem

$$\begin{array}{ll} \underset{\alpha_i}{\text{maximize}} & \displaystyle -\frac{1}{2}\sum_{i,j=1}^{m} \alpha_i\alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle + \sum_{i=1}^{m} \alpha_i \\ \text{subject to} & \alpha_i \in [0, C_i] \text{ for all } 1 \leq i \leq m \text{ and } \displaystyle\sum_{i=1}^{m} \alpha_i y_i = 0 \end{array}$$

4. The only difference to the standard SVM solution is that now $\alpha_i \in [0, C_i]$ rather than $[0, C]$. This means that we can now individually adjust the "force" each observation $(\mathbf{x}_i, y_i)$ applies.

Problem 3 (Perceptron Algorithm and Stochastic Gradient Descent, 15 Points)
We use a loss function $c(\mathbf{x}, y, f(\mathbf{x})) = |y - f(\mathbf{x})|_\varepsilon$ and want to minimize the **empirical risk** $R_{\text{emp}}[f]$.

1. *Compute derivatives of c with respect to $\mathbf{w}$, b for a linear model $f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b$.*
2. *State the general gradient descent rule regarding $\mathbf{w}, b$ for arbitrary c.*
3. *State the gradient descent rule for c specified as above.*
4. *With a learning rate $\eta = 0.1$ and $\varepsilon = 0.5$ calculate the values of $\mathbf{w}, b$ for the first four steps. The data set of (x, y) is given by $(1, 1), (3, 5), (2, 3), (1, 2)$ and we initialize $\mathbf{w} = \mathbf{0}$ and $b = 0$.*
5. *Modify the stochastic gradient algorithm using $|y - f(\mathbf{x})|_\varepsilon$ as a loss function such that it works in feature space using kernels, i.e. using $\Phi(\mathbf{x})$ and $k(\mathbf{x}, \mathbf{x}')$.*

Solution

1. The loss function $c(\mathbf{x}, y, f(\mathbf{x})) = |y - f(\mathbf{x})|_\varepsilon$ is linear for deviations larger than ε and zero otherwise. It is not differentiable at $\pm\varepsilon$ (but we ignore that below). We obtain

$$\begin{aligned} \partial_b |y - \langle \mathbf{w}, \mathbf{x} \rangle - b|_\varepsilon &= \begin{cases} 1 & \text{if } y + \varepsilon < \langle \mathbf{w}, \mathbf{x} \rangle + b \\ -1 & \text{if } y - \varepsilon > \langle \mathbf{w}, \mathbf{x} \rangle + b \\ 0 & \text{otherwise} \end{cases} \\ \partial_{\mathbf{w}} |y - \langle \mathbf{w}, \mathbf{x} \rangle - b|_\varepsilon &= \begin{cases} \mathbf{x} & \text{if } y + \varepsilon < \langle \mathbf{w}, \mathbf{x} \rangle + b \\ -\mathbf{x} & \text{if } y - \varepsilon > \langle \mathbf{w}, \mathbf{x} \rangle + b \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

2. For arbitrary c the stochastic gradient descent rule for a given learning rate Λ is

$$\begin{aligned} \mathbf{w} &\to \mathbf{w} - \Lambda \partial_{\mathbf{w}} c(\mathbf{x}, y, \langle \mathbf{w}, \mathbf{x} \rangle + b) \\ b &\to b - \Lambda \partial_b c(\mathbf{x}, y, \langle \mathbf{w}, \mathbf{x} \rangle + b) \end{aligned}$$

3. All we have to do is substitute the derivatives from 1. into the update equations given by 2.

$$b \to b + \Lambda \begin{cases} -1 & \text{if } y + \varepsilon < \langle \mathbf{w}, \mathbf{x} \rangle + b \\ 1 & \text{if } y - \varepsilon > \langle \mathbf{w}, \mathbf{x} \rangle + b \\ 0 & \text{otherwise} \end{cases} \quad \text{and } \mathbf{w} \to \mathbf{w} + \Lambda \begin{cases} -\mathbf{x} & \text{if } y + \varepsilon < \langle \mathbf{w}, \mathbf{x} \rangle + b \\ \mathbf{x} & \text{if } y - \varepsilon > \langle \mathbf{w}, \mathbf{x} \rangle + b \\ 0 & \text{otherwise} \end{cases}$$

4. We step through the data ($\eta = 0.1$) and obtain

   Step 1: $f(1) = 0$, however $y = 1$, hence update $\mathbf{w} = 0 + 0.1 \cdot 1, b = 0 + 0.1$.

   Step 2: $f(3) = 0.4$, however $y = 5$, hence update $\mathbf{w} = 0.1 + 0.1 \cdot 3, b = 0.1 + 0.1$.

   Step 3: $f(2) = 1$, however $y = 3$, hence update $\mathbf{w} = 0.4 + 0.1 \cdot 2, b = 0.2 + 0.1$.

   Step 4: $f(1) = 0.8$, however $y = 2$, hence update $\mathbf{w} = 0.6 + 0.1 \cdot 1 = 0.7, b = 0.3 + 0.1 = 0.4$.

5. All we have to do is replace $\mathbf{x}$ by $\Phi(\mathbf{x})$ wherever possible, and also keep track of $\mathbf{w} = \sum_{i=1}^t \alpha_i \Phi(\mathbf{x}_i)$ at step $t + 1$. Finally, $f(\mathbf{x}) = \sum_{i=1}^t \alpha_i k(\mathbf{x}_i, \mathbf{x})$. This yields

$$\begin{aligned} b &\to b + \Lambda \begin{cases} -1 & \text{if } y + \varepsilon < \sum_{i=1}^t \alpha_i k(\mathbf{x}_i, \mathbf{x}_{t+1}) + b \\ 1 & \text{if } y - \varepsilon > \sum_{i=1}^t \alpha_i k(\mathbf{x}_i, \mathbf{x}_{t+1}) + b \\ 0 & \text{otherwise} \end{cases} \\ \alpha_{t+1} &= \begin{cases} -\Lambda & \text{if } y + \varepsilon < \sum_{i=1}^t \alpha_i k(\mathbf{x}_i, \mathbf{x}_{t+1}) + b \\ \Lambda & \text{if } y - \varepsilon > \sum_{i=1}^t \alpha_i k(\mathbf{x}_i, \mathbf{x}_{t+1}) + b \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

   Here we keep the old α_i with $i \leq t$ and only uptdate the new α_{t+1}.

Problem 4 (Admissible Kernel, 15 Points)

1. *Show that the matrix $K_{ij} := y_i y_j k(\mathbf{x}_i, \mathbf{x}_j)$ is positive semidefinite if k satisfies Mercer's condition.*
2. *Show that if k_1, k_2 satisfy Mercer's condition, then also any $k = \alpha_1 k_1 + \alpha_2 k_2$ with $\alpha_1, \alpha_2 \geq 0$ also satisfy Mercer's condition.*
3. *Show that the kernel $k(\mathbf{x}, \mathbf{x}') = \kappa(\langle \mathbf{x}, \mathbf{x}' \rangle)$ with $\kappa(\xi) = \sum_{l=1}^{n} c_l \xi^l$ and $c_l \geq 0$ satisfies Mercer's condition.*
4. *Show that the kernel $k(x, x') = \frac{1}{1+(x-x')^2}$ with $x, x' \in \mathbb{R}$ satisfies Mercer's condition. Discuss its regularization properties.*

Solution

1. We know that the matrix $\tilde{K} \in \mathbb{R}^{m \times m}$ with $\tilde{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ is positive semidefinite, i.e. for any $\alpha \in \mathbb{R}^m$ we have $\alpha^\top \tilde{K} \alpha \geq 0$. This is in particular also the case for $\beta \in \mathbb{R}^m$ where $\beta_i = y_i \alpha_i$. Therefore

$$0 \leq \beta^\top \tilde{K} \beta = \sum_{i,j=1}^{m} \beta_i \beta_j \tilde{K}_{ij} = \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \tilde{K}_{ij} = \sum_{i,j=1}^{m} \alpha_i \alpha_j K_{ij} = \alpha^\top K \alpha$$

   In other words, K is positive semidefinite.

2. We exploit linearity of the integral and obtain for Mercer's condition on k

$$\begin{aligned} & \int\int k(x,x') f(x) f(x') dx dx' \\ = & \int\int (\alpha_1 k_1(x,x') + \alpha_2 k_2(x,x'))\, f(x) f(x') dx dx' \\ = & \alpha_1 \int\int k_1(x,x') f(x) f(x') dx dx' + \alpha_2 \int\int k_2(x,x') f(x) f(x') dx dx' \geq 0 \end{aligned}$$

   The last inequality follows since k_1, k_2 are both Mercer kernels and α_1, α_2 nonnegative.

3. We know that $k(\mathbf{x}, \mathbf{x}') = \langle \mathbf{x}, \mathbf{x}' \rangle^l$ is a Mercer kernel. Furthermore we know from 2. that also sums of Mercer kernels are Mercer kernels, provided the expansion coefficients are nonnegative. This means that the same must hold for the sum of several polynomial kernels as required in 3. and we have a Mercer kernel.

4. All we need to do is look at the Fourier transform of $\frac{1}{1+(x-x')^2}$. Knowing that (up to scaling factors) the Fourier transform of $\exp(-|x|)$ is $\frac{1}{1+\omega^2}$ (see lecture notes about the Laplacian kernel) and using the symmetry of k yields that the fourier transfrom of $\frac{1}{1+x^2}$ is, up to scaling factors, $\exp(-|\omega|)$.

   Therefore, the kernel of 4. is a Mercer kernel. Finally, we can see that the spectrum of k decays exponentially (cf. $\exp(-|\omega|)$) which means that it is smoother than the Laplacian but less smooth than the Gaussian kernel.

Problem 5 (Hilbert Spaces, 15 Points)
We begin with the quadratic form on functions $f : \mathbb{R} \to \mathbb{R}$.

$$q(f) := \|f\|^2 + 2\|f'\|^2 + 3\|f''\|^2 \tag{3}$$

1. *Use the polarization inequality to recover the dot product underlying the definition of the quadratic form $q(f)$, i.e. find a bilinear form from $q(f)$ such that $\langle f, f\rangle_{\mathcal{H}} = q(f)$. Show that what you obtained is a dot product.*
2. *Compute $q(f)$ for $f = c_0 + \sum_{l=1}^{n}(s_l \sin(lx) + c_l \cos(lx))$.*
3. *Compute the representation of $q(f)$ in the Fourier domain, i.e. in terms of $\tilde{f}(\omega)$.*
4. *Show that the corresponding kernel in feature space setting is translation invariant, i.e. $k(x, x') = k(x - x')$.* **Note: you need not compute k for that purpose!**

Solution

1. The polarization reconstructs a dot product via
$$\frac{1}{4}\left(q(f+g) - q(f-g)\right) = \langle f, g\rangle.$$
Using this for the quadratic form above and using that $\|f\|^2 = \langle f, f\rangle$ we obtain
$$\langle f, g\rangle_{\mathcal{H}} = \langle f, g\rangle + 2\langle f', g'\rangle + 3\langle f'', g''\rangle.$$
Substituting f in place of g recovers the quadratic form we started with. To show that $\langle \cdot, \cdot\rangle_{\mathcal{H}}$ is a dot product we need to show bilinearity of the overall expression. Instead of writing out the equations we simply exploit that $\langle \cdot, \cdot\rangle$ is a dot product and that f' is obtained by a linear operation from f (and f'', g', g'' analogously). This completes the proof.

2. Before we begin, note that $1, \sin lx$, and $\cos lx$ are orthogonal with respect to the dot product if we choose a suitable domain. Since the domain of integration is not specified, we pick one, namely $[0, 2\pi]$. Therefore we have
$$q(f) = 2\pi c_0 + \pi \sum_{l=1}^{n}(s_l^2 + c_l^2) + 2\pi \sum_{l=1}^{n} l^2(s_l^2 + c_l^2) + 3\pi \sum_{l=1}^{n} l^4(s_l^2 + c_l^2).$$

3. Computing $q(f)$ in Fourier domain means using the fact that for $\tilde{f}$ being the Fourier transform of f we have $\tilde{f}'(\omega) = \omega\tilde{f}$, $\tilde{f}''(\omega) = \omega^2\tilde{f}$, and $\langle f, f\rangle = \langle f', f'\rangle$. Therefore we may write $q(f)$ as
$$q(f) = \|\tilde{f}\|^2 + 2\|\tilde{f}'\|^2 + 3\|\tilde{f}''\|^2 = \|\tilde{f}\|^2 + 2\|\omega\tilde{f}\|^2 + 3\|\omega^2\tilde{f}\|^2 = \int \tilde{f}^2(\omega)(1 + 2\omega^2 + 3\omega^4)d\omega$$
This is the Fourier representation of $q(f)$.

4. From the lectures we know that a regularization term of the form $\|f\|^2_{\mathcal{H}} := \int P^{-1}(\omega)\tilde{f}^2(\omega)d\omega$ corresponds to a tranlation invariant kernel. In our case $P(\omega) = \frac{1}{1+2\omega^2+3\omega^4}$ and therefore k is the Fourier (back-)transform of $P(\omega)$.