CO-496: Mathematics for Inference and Machine Learning

Problem Sheet for Tutorial 6

Problem 1
Whitened SVMs. Assume the centralised data $\mathbf{x}_i$, $i = 1, \dots, n$. Assume further that each datum comes with a label $y_i = -1, 1$. Let $\mathbf{S}_t$ is the covariance matrix of the data and D a positive constant between 0 and 1. Then, formulate the dual of the following whitened SVM optimisation problem

$$\begin{aligned} &\min_{\mathbf{w},b,\xi_i} \frac{1-D}{2}\mathbf{w}^T\mathbf{w} + \frac{D}{2}\mathbf{w}^T\mathbf{S}_t\mathbf{w} + C\sum_{i=1}^{n}\xi_i \\ &\text{s.t.} y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 1 - \xi_i, \xi_i \geq 0. \end{aligned} \tag{1}$$

Problem 2
Kernel Discriminant Analysis.
Assume the data samples $\mathbf{x}_1, \dots, \mathbf{x}_n$, split in C different classes. Assume you are given a positive definite kernel k which defines an implicit Hilbert space on the vectors $\phi(\mathbf{x}_i) \in \mathcal{H}$. Assume the Kernel Discriminant Analysis (KDA) optimisation problem

$$\begin{aligned} \max_{\mathbf{W}_\Phi} \; & \operatorname{tr}(\mathbf{W}_\Phi^T\mathbf{S}_b^\Phi\mathbf{W}_\Phi) \\ \text{subject to} \; & \mathbf{W}_\Phi^T\mathbf{S}_w^\Phi\mathbf{W}_\Phi = \mathbf{I} \end{aligned} \tag{2}$$

where $\mathbf{S}_b^\Phi$ and $\mathbf{S}_w^\Phi$ are the between and within class scatter matrices, respectively, defined in the Hilbert space of the vectors $\phi(\mathbf{x}_i) \in \mathcal{H}$. Find the optimal $\mathbf{W}_\Phi$ and extract the features from a test vector $\phi(\mathbf{y})$.

Problem 3
SVMs are very powerful methods for classification. Nevertheless, they are sensitive to affine transformations of the data and to directions with large data spread. Maximum margin solutions may be misled by the spread of data and preferentially separate classes along large spread directions. An alternative is the Relative Margin Machines (RMM) which creates a relative margin, controlled by a parameter $B > 1$. Assume the centralised data $\mathbf{x}_i$, $i = 1, \dots, n$. Assume further that each datum comes with a label $y_i = -1, 1$. Then, the RMM optimisation problem

$$\begin{aligned}
&\min_{\mathbf{w},b,\xi_i} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n}\xi_i \qquad (3)\\
&\text{s.t. } y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 1 - \xi_i, \xi_i \geq 0\\
&\quad \mathbf{w}^T\mathbf{x}_i + b \leq B\\
&\quad -\mathbf{w}^T\mathbf{x}_i - b \leq B.
\end{aligned}$$

Formulate the dual of the above optimisation problem.

Solutions

Problem 1 Firstly, we formulate the Lagrangian.

$$\mathcal{L}(\theta) = \frac{1-D}{2}\|\mathbf{w}\,\|^2 + \frac{D}{2}\mathbf{w}^T\mathbf{S}_t\mathbf{w} + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}a_i\left[y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) - 1 + \xi_i\right] - \sum_{i=1}^{n}\gamma_i\xi_i, \tag{4}$$

where $\theta = \{\mathbf{w}, b, \xi_i, a_i, \gamma_i\}$. We then take partials of the Lagrangian w.r.t. $\mathbf{w}, b, xi_i$ and set them equal to 0. We then have

$$\begin{aligned} \frac{\partial\mathcal{L}(\theta)}{\partial\mathbf{w}} &= 0 \Rightarrow \\ (1-D)\mathbf{w} + D\mathbf{S}_t\mathbf{w} &= \sum_{i=1}^{n}a_iy_i\mathbf{x}_i \Rightarrow \\ \mathbf{w} &= [(1-D)\mathbf{I} + D\mathbf{S}_t]^{-1}\cdot\sum_{i=1}^{n}a_iy_i\mathbf{x}_i, \end{aligned} \tag{5}$$

$$\begin{aligned} \frac{\partial\mathcal{L}(\theta)}{\partial b} &= 0 \Rightarrow \\ \mathbf{a}^T\mathbf{y} &= 0, \end{aligned} \tag{6}$$

$$\begin{aligned} \frac{\partial\mathcal{L}(\theta)}{\partial\xi_i} &= 0 \Rightarrow \\ 0 \leq &a_i \leq C. \end{aligned} \tag{7}$$

Re-writing $\mathbf{w} = \sum_{i=1}^{n}a_iy_i\mathbf{A}\mathbf{x}_i$, where $\mathbf{A} = [(1-D)\mathbf{I} + D\mathbf{S}_t]^{-1}$, and utilizing (6), (7), we get

$$\begin{aligned} &\max_{\mathbf{a}}\mathbf{1}^T\mathbf{a} - \frac{1}{2}\mathbf{a}^T\mathbf{K}^{\mathbf{S}_t}\mathbf{a} \\ \text{s.t.}\quad &\mathbf{y}^T\mathbf{a} = 0 \\ &0 \leq a_i \leq C, \end{aligned} \tag{8}$$

where $\mathbf{K}^{\mathbf{S}_t} = \left[y_iy_j\mathbf{x}_i^T\mathbf{A}\mathbf{x}_y\right]$.

Problem 2
We assume that $\mathbf{W}_\Phi = \mathbf{U}_\Phi\mathbf{V}$ and we want some $\mathbf{U}_\Phi$ that diagonalises $\mathbf{S}_w^\Phi$, i.e., $\mathbf{U}_\Phi^T\mathbf{S}_w^\Phi\mathbf{U}_\Phi = \mathbf{I}$. We cannot apply eigen-analysis on

$$\mathbf{S}_w^\Phi = \bar{\mathbf{X}}_\Phi(\mathbf{I} - \mathbf{E})\bar{\mathbf{X}}_\Phi^T \tag{9}$$

where $\bar{\mathbf{X}}_\Phi = [\phi(\mathbf{x}_1) - \mathbf{m}_\Phi, \dots, \phi(\mathbf{x}_n) - \mathbf{m}_\Phi] = [\phi(\mathbf{x}_1), \dots, \phi(\mathbf{x}_n)](\mathbf{I} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T)$ is the centralised data matrix in the feature space and $\mathbf{1}_n$ is a vector of n ones. Moreover, $\mathbf{E} = \text{diag}\{\frac{1}{n_1}\mathbf{1}_{n_1}\mathbf{1}_{n_1}^T, \dots, \frac{1}{n_C}\mathbf{1}_{n_C}\mathbf{1}_{n_C}^T\}$. Since $\mathbf{I} - \mathbf{E}$ is an idempotent matrix, in order to compute the eigenvalues of $\mathbf{S}_w^\Phi$, we perform eigen-analysis on

$$\mathbf{K}_w = (\mathbf{I} - \mathbf{E})\bar{\mathbf{X}}_\Phi^T\bar{\mathbf{X}}_\Phi(\mathbf{I} - \mathbf{E}) = (\mathbf{I} - \mathbf{E})\bar{\mathbf{K}}(\mathbf{I} - \mathbf{E}) = \mathbf{V}_w\mathbf{\Lambda}_w\mathbf{V}_w^T \tag{10}$$

where $\bar{\mathbf{K}}$ is the centralised kernel matrix. Therefore, $\mathbf{U}_\Phi$ is given by

$$\mathbf{U}_\Phi = \bar{\mathbf{X}}_\Phi(\mathbf{I} - \mathbf{E})\mathbf{V}_w\mathbf{\Lambda}_w^{-1}. \tag{11}$$

Now we project the between class scatter matrix $\mathbf{S}_b^\Phi = \bar{\mathbf{X}}_\Phi\mathbf{E}\bar{\mathbf{X}}_\Phi^T$ as

$$\begin{aligned}\tilde{\mathbf{S}}_b = \mathbf{U}_\Phi^T\mathbf{S}_b^\Phi\mathbf{U}_\Phi &= \mathbf{\Lambda}_w^{-1}\mathbf{V}_w^T(\mathbf{I} - \mathbf{E})\bar{\mathbf{X}}_\Phi^T\bar{\mathbf{X}}_\Phi\mathbf{E}\bar{\mathbf{X}}_\Phi^T\bar{\mathbf{X}}_\Phi(\mathbf{I} - \mathbf{E})\mathbf{V}_w\mathbf{\Lambda}_w^{-1} \\ &= \mathbf{\Lambda}_w^{-1}\mathbf{V}_w^T(\mathbf{I} - \mathbf{E})\bar{\mathbf{K}}\mathbf{E}\bar{\mathbf{K}}(\mathbf{I} - \mathbf{E})\mathbf{V}_w\mathbf{\Lambda}_w^{-1}\end{aligned} \tag{12}$$

which can be computed using only the kernel. As a result, the optimisation problem is transformed to

$$\begin{aligned}\max_{\mathbf{V}} \; & \text{tr}(\mathbf{V}^T\tilde{\mathbf{S}}_b\mathbf{V}) \\ \text{subject to } \; & \mathbf{V}^T\mathbf{V} = \mathbf{I}\end{aligned} \tag{13}$$

which is solved by selecting as columns of $\mathbf{V}$ the $C - 1$ eigenvectors of $\tilde{\mathbf{S}}_b$ that correspond to non-zero eigenvalues.

We can extract features from a test sample $\phi(\mathbf{y})$ as

$$\mathbf{W}_\Phi^T\phi(\mathbf{y}) = \mathbf{V}^T\mathbf{\Lambda}_w^{-1}\mathbf{V}_w^T(\mathbf{I} - \mathbf{E})(\mathbf{I} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T)\mathbf{X}_\Phi^T\phi(\mathbf{y}) \tag{14}$$

and $\mathbf{X}_\Phi^T\phi(\mathbf{y})$ can be computed as $\mathbf{X}_\Phi^T\phi(\mathbf{y}) = [k(\mathbf{x}_1, \mathbf{y}), \dots, k(\mathbf{x}_n, \mathbf{y})]$.

Problem 3

Firstly, we formulate the Lagrangian.

$$\mathcal{L}(\theta) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^n \xi_i - \sum_{i=1}^n a_i\left[y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) - 1 + \xi_i\right] + \tag{15}$$

$$+\sum_{i=1}^n \gamma_i\left(\mathbf{w}^T\mathbf{x}_i + b - B\right) + \sum_{i=1}^n k_i\left(-\mathbf{w}^T\mathbf{x}_i - b - B\right) - \sum_{i=1}^n \lambda_i\xi_i, \tag{16}$$

where $\theta = \{\mathbf{w}, b, \xi_i, a_i, \gamma_i, k_i, \lambda_i\}$.

We then take partials of the Lagrangian w.r.t. $\mathbf{w}, b, xi_i$ and set them equal to 0. We then have

$$\begin{aligned}\frac{\partial \mathcal{L}(\theta)}{\partial \mathbf{w}} &= 0 \Rightarrow \\ \mathbf{w} &= \sum_{i=1}^{n} (a_i y_i - \gamma_i + k_i)\mathbf{x}_i \Rightarrow \end{aligned} \tag{17}$$

$$\begin{aligned}\frac{\partial \mathcal{L}(\theta)}{\partial b} &= 0 \Rightarrow \\ \sum_{i=1}^{n} a_i y_i + \sum_{i=1}^{n} \gamma_i - \sum_{i=1}^{n} k_i &= 0, \Rightarrow \\ \mathbf{y}^T\mathbf{a} - \mathbf{1}^T\boldsymbol{\gamma} + \mathbf{1}^T\mathbf{k} &= 0 \end{aligned} \tag{18}$$

$$\begin{aligned}\frac{\partial \mathcal{L}(\theta)}{\partial \xi_i} &= 0 \Rightarrow \\ 0 \leq & a_i \leq C. \end{aligned} \tag{19}$$

Re-writing $\mathbf{w} = \sum_{i=1}^{n} (a_i y_i - \gamma_i + k_i)\mathbf{x}_i$ and utilizing (18), (19), we get

$$\max_{\mathbf{a},\boldsymbol{\gamma},\boldsymbol{k}} -\frac{1}{2} \left(\mathbf{a} \odot \mathbf{y} - \boldsymbol{\gamma} + \mathbf{k}\right)^T \mathbf{X}^T\mathbf{X} \left(\mathbf{a} \odot \mathbf{y} - \boldsymbol{\gamma} + \mathbf{k}\right) + \mathbf{a}^T\mathbf{1} - B\boldsymbol{\gamma}^T\mathbf{1} - B\mathbf{k}^T\mathbf{1}, \tag{20}$$

where $\mathbf{a} \odot \mathbf{b}$ denotes the Hadamard (element-wise) product between two vectors, i.e., $(\mathbf{ab})_i = \mathbf{a}_i\mathbf{b}_i$.