SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Homework #6

Course: *Machine Learning (CS405)* – Professor: *Qi Hao*
Due date: *October 7th, 2019*

Homework Submission Instructions. Please write up your responses to the following problems clearly and concisely. We require you to write up your responses with A4 paper. You are allowed and encouraged to work together. You may discuss the homework to understand the problem and reach a solution in groups. However, *each student must write down the solution independently*. You must understand the solution well enough in order to reconstruct it by yourself. (This is for your own benefit: you have to take the exams alone.)

- **Written Homeworks.** All calculation problems **MUST** be written on single-sided A4 paper. You should bring and hand in it before class on the day of the deadline. Submitting the scan or photo version on Sakai will **NOT** be accepted.
- **Coding Homeworks.** All coding assignments will be done in Jupyter Notebooks. We will provide a `.ipynb` template for each assignment. Your final submission will be a `.ipynb` file with your answers and explanations (you should know how to write in Markdown or LaTeX). Make sure that all packages you need are imported at the beginning of the program, and your `.ipynb` file should **work step-by-step without any error**.

Question 1

Suppose we have a data set of input vectors $\{\mathbf{x}_n\}$ with corresponding target values $t_n \in \{-1, 1\}$, and suppose that we model the density of input vectors within each class separately using a Parzen kernel density estimator with a kernel $k(\mathbf{x}, \mathbf{x}')$.

(a) Write down the minimum misclassification-rate decision rule assuming the two classes have equal prior probability.

(b) Show that if the kernel is chosen to be $k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^{\mathrm{T}}\mathbf{x}'$, then the classification rule reduces to simply assigning a new input vector to the class having the closest mean.

(c) Show that if the kernel takes the form $k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^T\phi(\mathbf{x}')$, then the classification is based on the closest mean in the feature space $\phi(\mathbf{x})$.

Answer. From Bayes' theorem we have

$$p(t|\mathbf{x}) \propto p(\mathbf{x}|t)p(t)$$

where

$$p(\mathbf{x}|t) = \frac{1}{N_t} \sum_{n=1}^{N} \frac{1}{Z_k} k(\mathbf{x}, \mathbf{x}_n) \delta(t, t_n).$$

Here N_t is the number of input vectors with label t(+1 or -1) and $N = N_{+1} + N_{-1}$. $\delta(t, t_n)$ equals 1 if $t = t_n$ and 0 otherwise. Z_k is the noemalisation constant for the kernel. The minimum misclassification-rate is archieved if, for each new input vector, $\tilde{\mathbf{x}}$, we chose $\tilde{t}$ to maximise $p(\tilde{t}|\tilde{\mathbf{x}})$. With equal class priors, this is equivalent to maximizing $p(\tilde{\mathbf{x}}|\tilde{t})$ and thus

$$\tilde{t} = \begin{cases} +1 & \text{iff } \frac{1}{N_{+1}} \sum_{i:t_i=+1} k(\tilde{\mathbf{x}}, \mathbf{x}_i) \geq \frac{1}{N_{-1}} \sum_{j:t_j=-1} k(\tilde{\mathbf{x}}, \mathbf{x}_j) \\ -1 & \textit{otherwise}. \end{cases}$$

Here we have dropped the factor $1/Z_k$ since it only acts as a commom scaling factor. Using the encoding scheme for the label, this classification rule can be written in the more compact form

$$\tilde{t} = \text{sign}(\sum_{n=1}^{N} \frac{t_n}{N_{t_n}} k(\tilde{\mathbf{x}}, \mathbf{x}_n)).$$

Now we take $k(\tilde{\mathbf{x}}, \mathbf{x}_n) = \mathbf{x^T x_n}$, which results in the kernel density

$$p(\mathbf{x}|t = +1) = \frac{1}{N_{+1}} \sum_{n:t_n=+1} \mathbf{x^T} \mathbf{x}_n = \mathbf{x^T} \bar{\mathbf{x}}^+.$$

Here, the sum in the middle experssion runs overall vectors $\mathbf{x}_n$ for which $t_n = +1$ and $\tilde{\mathbf{x}}^+$ denotes the mean of these vectors, with the corresponding definition for the negative class. Note that this density is improper, since it cannot be normalized. However, we can still compare likelihoods under this density, resulting in the claasification rule

$$\tilde{t} = \begin{cases} +1 & \text{if } \tilde{\mathbf{x}}^T \tilde{\mathbf{x}}^+ \geq \tilde{\mathbf{x}}^T \tilde{\mathbf{x}}^- \\ -1 & \textit{otherwise}. \end{cases}$$

The same argument would of course also apply in the feature space $\phi(\mathbf{x})$.

Question 2

Consider the logistic regression model with a target variable $t \in \{-1, 1\}$. If we define $p(t = 1|y) = \sigma(y)$ where $y(\mathbf{x})$ is given by

$$y(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) + b,$$

show that the negative log likelihood, with the addition of a quardratic regularization term, takes the form

$$\sum_{n=1}^{N} E_{LR}(y_n t_n) + \lambda \|\mathbf{w}\|^2.$$

Answer. If $p(t=1|y)=\sigma(y)$, then

$$p(t=-1|y)=1-p(t=1|y)=1-\sigma(y)=\sigma(-y).$$

Thus, given i.i.d. data $\mathcal{D}=\{(t_1,\mathbf{x}_n),...,(t_N,\mathbf{x}_N)\}$, we can write the correspongding likelihood as

$$p(\mathcal{D})=\prod_{t_n=1}\sigma(y_n)\prod_{t_{n'}=-1}\sigma(-y_{n'})=\prod_{n=1}^{N}\sigma(t_n y_n),$$

where $y_n=y(\mathbf{x}_n)$. Taking the negative logarithm of this, we get

$$\begin{aligned}-\ln p(\mathcal{D}) &= -\ln\prod_{n=1}^{N}\sigma(t_n y_n)\\ &= \sum_{n=1}^{N}\ln\sigma(t_n y_n)\\ &= \sum_{n=1}^{N}\ln(1+\exp(-t_n y_n)),\end{aligned}$$

Combining this with the regularization term $\lambda\|\mathbf{w}\|^2$, we obtain

$$\sum_{n=1}^{N}E_{LR}(y_n t_n)+\lambda\|\mathbf{w}\|^2.$$

Question 3

By performing the Guassian integral over $\mathbf{w}$ in

$$p(\mathbf{t}|\mathbf{X},\alpha,\beta)=\int p(\mathbf{t}|\mathbf{X},\mathbf{w},\beta)p(\mathbf{w},\alpha)\mathrm{d}\mathbf{w}$$

using the technique of complrting the square in the exponential, derive the result for the marginal likelihood function in the regression RVM:

$$\ln p(\mathbf{t}|\mathbf{X},\alpha,\beta)=-\frac{1}{2}\{N\ln(2\pi)+\ln|\mathbf{C}|+\mathbf{t}^{\mathbf{T}}\mathbf{C}^{-\mathbf{1}}\mathbf{t}\}$$

Answer.

$$p(\mathbf{t}|\mathbf{X},\alpha,\beta)=(\frac{\beta}{2\pi})^{N/2}\frac{1}{(2\pi)^{N/2}}\prod_{i=1}^{M}\alpha_i\int\exp\{-E(\mathbf{w})\}\mathrm{d}\mathbf{w}$$

where

$$E(\mathbf{w})=\frac{\beta}{2}\|\mathbf{t}-\mathbf{\Phi}\mathbf{w}\|^2+\frac{1}{2}\mathbf{w}^{\mathbf{T}}\mathbf{A}\mathbf{w},\ \text{and}\ \mathbf{A}=\mathrm{diag}(\alpha).$$

Completing the square over $\mathbf{w}$, we get

$$E(\mathbf{w})=\frac{1}{2}(\mathbf{w}-\mathbf{m})^{\mathbf{T}}\mathbf{\Sigma}^{-\mathbf{1}}(\mathbf{w}-\mathbf{m})+E(\mathbf{t})$$

where $\mathbf{m}$ and $\mathbf{\Sigma}$ are given by

$$\mathbf{m}=\beta\mathbf{\Sigma}\mathbf{\Phi}^{\mathbf{T}}\mathbf{t},\qquad \mathbf{\Sigma}=(\mathbf{A}+\beta\mathbf{\Phi}^{\mathbf{T}}\mathbf{\Phi})^{-1},$$

and

$$E(\mathbf{t}) = \frac{1}{2}(\beta \mathbf{t}^{\mathbf{T}}\mathbf{t} - \mathbf{m}^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\mathbf{m}).$$

Using $E(\mathbf{w})$, we can evaluate the integral in $E(\mathbf{t})$ to obtain

$$\int \exp\{-E(\mathbf{w})\}\mathrm{d}\mathbf{w} = \exp\{-E(\mathbf{t})\}(2\pi)^{M/2}|\boldsymbol{\Sigma}|^{1/2}.$$

Considering this as a function of $\mathbf{t}$, that we only need to deal with the factor $exp\{-E(\mathbf{t})\}$. Then we can re-write $E(\mathbf{t})$ as follows

$$\begin{aligned}
E(\mathbf{t}) &= \frac{1}{2}(\beta \mathbf{t}^{\mathbf{T}}\mathbf{t} - \mathbf{m}^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\mathbf{m}) \\
&= \frac{1}{2}(\beta \mathbf{t}^{\mathbf{T}}\mathbf{t} - \beta \mathbf{t}^{\mathbf{T}}\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-\mathbf{1}}\boldsymbol{\Sigma}\boldsymbol{\Phi}^{\mathbf{T}}\mathbf{t}\beta) \\
&= \frac{1}{2}\mathbf{t}^{T}(\beta\mathbf{I} - \beta\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^{T}\beta)\mathbf{t} \\
&= \frac{1}{2}\mathbf{t}^{T}(\beta\mathbf{I} - \beta\boldsymbol{\Phi}(\mathbf{A} + \beta\boldsymbol{\Phi}^{\mathbf{T}}\boldsymbol{\Phi})^{-\mathbf{1}}\boldsymbol{\Phi}^{\mathbf{T}}\beta)\mathbf{t} \\
&= \frac{1}{2}\mathbf{t}^{T}(\beta^{-1}\mathbf{I} + \boldsymbol{\Phi}\mathbf{A}^{-\mathbf{1}}\boldsymbol{\Phi}^{\mathbf{T}})^{-\mathbf{1}}\mathbf{t} \\
&= \frac{1}{2}\mathbf{t}^{\mathbf{T}}\mathbf{C}^{-\mathbf{1}}\mathbf{t}.
\end{aligned}$$

This gives us the last term on the r.h.s. of the target equation; the two preceding terms are given implicitly, as they form the normalization constant for the posterior Gaussian diatribution $p(\mathbf{t}|\mathbf{X},\alpha,\beta)$.

Question 4

Show that direct maximization of the log marginal likelihood

$$\begin{aligned}
\ln p(\mathbf{t}|\mathbf{X},\alpha,\beta) &= \ln \mathcal{N}(\mathbf{t}|0,\mathbf{C}) \\
&= -\frac{1}{2}\{N\ln(2\pi) + \ln|\mathbf{C}| + \mathbf{t}^{\mathbf{T}}\mathbf{C}^{-\mathbf{1}}\mathbf{t}\}
\end{aligned}$$

for the regression relevance vector machine leads to the re-estimation equations

$$\alpha_i^{new} = \frac{\gamma_i}{m_i^2}$$

and

$$(\beta^{new})^{-1} = \frac{||\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}||^2}{N - \sum_i \gamma_i}$$

where γ_i is defined by

$$\gamma_i = 1 - \alpha_i \Sigma_{ii}$$

Answer. Using the results from Question 3, we can write $\ln p(\mathbf{t}|\mathbf{X},\alpha,\beta)$ in the form of :

$$\ln p(\mathbf{t}|\mathbf{X},\alpha,\beta) = \frac{N}{2}\ln\beta + \frac{1}{2}\sum_i^N \alpha_i - E(\mathbf{t}) - \frac{1}{2}\ln|\Sigma| - \frac{N}{2}\ln(2\pi)$$

By making use of $E(\mathbf{t}) = \frac{1}{2}(\beta \mathbf{t}^\mathbf{T}\mathbf{t} - \mathbf{m}^\mathbf{T}\boldsymbol{\Sigma}^{-1}\mathbf{m})$ and $\boldsymbol{\Sigma} = (\mathbf{A} + \beta\boldsymbol{\Phi}^\mathbf{T}\boldsymbol{\Phi})^{-1}$ together with $\frac{\partial}{\partial x}\ln|A| = \mathrm{Tr}(\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial x})$, we can take the derivatives of this w.r.t α_i, yielding

$$\frac{\partial}{\partial \alpha_i}\ln p(\mathbf{t}|\mathbf{X}, \alpha, \beta) = \frac{1}{2\alpha_i} - \frac{1}{2}\Sigma_{ii} - \frac{1}{2}m_i^2$$

Setting this to zero and re-arranging, we obtain

$$\alpha_i = \frac{1 - \alpha_i\Sigma_{ii}}{m_i^2} = \frac{\gamma_i}{m_i^2}$$

where we have used

$$\gamma_i = 1 - \alpha_i\Sigma_{ii}$$

Similarly, for β we see that

$$\frac{\partial}{\partial \beta}\ln p(\mathbf{t}|\mathbf{X}, \alpha, \beta) = \frac{1}{2}\left(\frac{N}{\beta} - ||\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}||^2 - \mathrm{Tr}[\boldsymbol{\Sigma}\boldsymbol{\Phi}^\mathbf{T}\boldsymbol{\Phi}]\right)$$

Using $\boldsymbol{\Sigma}\Sigma = (A + \beta\boldsymbol{\Phi}^\mathbf{T}\boldsymbol{\Phi})^{-1}$, we can rewrite the argument of the trace operator as

$$\begin{aligned}
\boldsymbol{\Sigma}\boldsymbol{\Phi}^\mathbf{T}\boldsymbol{\Phi} &= \boldsymbol{\Sigma}\boldsymbol{\Phi}^\mathbf{T}\boldsymbol{\Phi} + \beta^{-1}\boldsymbol{\Sigma}\mathbf{A} - \beta^{-1}\boldsymbol{\Sigma}\mathbf{A} \\
&= \boldsymbol{\Sigma}(\boldsymbol{\Phi}^\mathbf{T}\boldsymbol{\Phi}\beta + \mathbf{A})\beta^{-1} - \beta^{-1}\boldsymbol{\Sigma}\mathbf{A} \\
&= (\mathbf{A} + \beta\boldsymbol{\Phi}^\mathbf{T}\boldsymbol{\Phi})^{-1}(\boldsymbol{\Phi}^\mathbf{T}\boldsymbol{\Phi}\beta + \mathbf{A})\beta^{-1} - \beta^{-1}\boldsymbol{\Sigma}\mathbf{A} \\
&= (\mathbf{I} - \mathbf{A}\boldsymbol{\Sigma})\beta^{-1}
\end{aligned}$$

Here the first factor on the r.h.s. of the last line equals $\gamma_i = 1 - \alpha_i\Sigma_{ii}$ written in matrix form. We can use this to set $\frac{\partial}{\partial \beta}\ln p(\mathbf{t}|\mathbf{X}, \alpha, \beta)$ equal to zero and then re-arrange to obtain $(\beta^{new})^{-1} = \frac{||\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}||^2}{N - \sum_i \gamma_i}$.

Question 5

Kernel functions implicitly define some mapping function $\phi(\cdot)$ that transforms an input instance $x \in \mathbb{R}^d$ to high dimensional space Q by giving the form of dot product in Q : $K(\mathbf{x}_i, \mathbf{x}_j) \equiv < \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) >$

(a) Prove that the kernel is symmetric, i.e. $K(\mathbf{x}_i, \mathbf{x}_j) = K(\mathbf{x}_j, \mathbf{x}_i)$

(b) Assume we use radial basis kernel function $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{1}{2}||\mathbf{x}_i - \mathbf{x}_j||^2)$. Thus there is some implicit unknown mapping function $\phi(x)$. Prove that for any two input instances $\mathbf{x}_i$ and $\mathbf{x}_j$, the squared Euclidean distance of their corresponding points in the feature space Q is less than 2, i.e. prove that $||\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)||^2 \leq 2$

Hint.

$$||\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)||^2 = < \phi(\mathbf{x}_i), \phi(\mathbf{x}_i) > + < \phi(\mathbf{x}_j), \phi(\mathbf{x}_j) > -2 \cdot < \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) >$$

Answer. (a) We have

$$K(\mathbf{x}_i, \mathbf{x}_j) =< \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) >=< \phi(\mathbf{x}_j), \phi(\mathbf{x}_i) >= K(\mathbf{x}_j, \mathbf{x}_i)$$

(b)

$$\begin{aligned} ||\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)||^2 &=< \phi(\mathbf{x}_i), \phi(\mathbf{x}_i) > + < \phi(\mathbf{x}_j), \phi(\mathbf{x}_j) > -2 \cdot < \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) > \\ &= K(\mathbf{x}_i, \mathbf{x}_i) + K(\mathbf{x}_j, \mathbf{x}_j) - 2 \cdot K(\mathbf{x}_i, \mathbf{x}_j) \\ &= 1 + 1 - 2\exp(-\frac{1}{2}||\mathbf{x}_i - \mathbf{x}_j||^2) \\ &< 2 \end{aligned}$$

Question 6

With the help of a kernel function, SVM attempts to construct a hyper-plane in the feature space Q that maximizes the margin between two classes. The classification decision of any $\mathbf{x}$ is made on the basis of the sign of

$$< \hat{\mathbf{w}}, \phi(x) > + \hat{w}_0 = \sum_{i \in SV} y_i \alpha_i K(\mathbf{x}_i, \mathbf{x}) + \hat{w}_0 = f(\mathbf{x}; \alpha, \hat{w}_0)$$

where $\hat{\mathbf{w}}$ and $\hat{w}_0$ are parameters for the classification hyper-plane in the feature space Q, SV is the set of support vectors, and α_i is the coefficient for the i-th support vector. Again we use the radial basis kernel function. Assume that the training instances are linearly separable in the feature space Q, and assume that the SVM finds a margin that perfectly separates the points.

If we choose a test point $\mathbf{x}_{far}$ which is far away from any training instance $\mathbf{x}_i$ (distance here is measured in the original space $\mathbb{R}^d$), prove that

$$f(\mathbf{x}_{far}; \alpha, \hat{w}_0) \approx \hat{w}_0$$

Hint. We have that

$$||\mathbf{x}_{far} - \mathbf{x}_i|| >> 0 \qquad \forall i \in SV$$

Answer.

$$\begin{aligned} & ||\mathbf{x}_{far} - \mathbf{x}_i|| >> 0 \qquad \forall i \in SV \\ \Rightarrow\ & K(\mathbf{x}_{far}, \mathbf{x}_i) \approx 0 \qquad \forall i \in SV \\ \Rightarrow\ & \sum_{i \in SV} y_i \alpha_i K(\mathbf{x}_i, \mathbf{x}) \approx 0 \\ \Rightarrow\ & f(\mathbf{x}_{far}; \alpha, \hat{w}_0) \approx \hat{w}_0 \end{aligned}$$