COS 598 – Machine Learning - Homework #4

Due: 5:00PM 04/09/2021

Homework Submission: Homeworks must be submitted via Brightspace as pdf files. This includes your code when appropriate. Please use a high quality scanner if possible, as found at the library or your departmental copy room. If you must use your phone, please don't just take photos (if possible), at least use an app like CamScanner that provides some correction for shading and projective transformations.

1) **Kernel Ridge Regression (25 pts).**

Recall that the error function for ridge regression (linear regression with L2 regularization) is:

$$E(\mathbf{w}) = (\Phi\mathbf{w} - \mathbf{t})^T(\Phi\mathbf{w} - \mathbf{t}) + \lambda\mathbf{w}^T\mathbf{w}$$

and its closed-form solution and model are:

$$\hat{\mathbf{w}} = (\Phi^T\Phi + \lambda I)^{-1}\Phi^T\mathbf{t} \text{ and } \hat{f}(\mathbf{x}) = \hat{\mathbf{w}}^T\phi(\mathbf{x}) = \mathbf{t}^T\Phi(\Phi^T\Phi + \lambda I)^{-1}\phi(\mathbf{x})$$

Now we want to kernelize ridge regression and allow non-linear models. Use the following matrix inverse lemma to derive the closed-form solution and model for kernelized ridge regression:

$$(P + QRS)^{-1} = P^{-1} - P^{-1}Q(R^{-1} + SP^{-1}Q)^{-1}SP^{-1}$$

where P is an $n \times n$ invertible matrix, R is a $k \times k$ invertible matrix, Q is an $n \times k$ matrix and S is a $k \times n$ matrix. In Lecture 14 we have shown the kernelized model only depends on the feature vectors $\phi(\mathbf{x})$ through inner products with other feature vectors and we provided the kernel ridge regression model. In this problem you implement KRR to a real datset.

(a) Apply kernelized ridge regression to the automobile mpg dataset. The training data and test data are provided in `auto_mpg_train.csv` and `auto_mpg_test.csv`, respectively. The first column is the mpg data while the other 7 columns are the features:

```
1. mpg: continuous
2. cylinders: multi-valued discrete
3. displacement: continuous
4. horsepower: continuous
5. weight: continuous
6. acceleration: continuous
7. model year: multi-valued discrete
8. origin: multi-valued discrete
```

We have normalized the feature data to the range [0,1]. Please apply the kernelized ridge regression to this dataset (use mpg as the target, the other 7 columns as features). Report the RMSE (Root Mean Square Error) of the models on the test data. Try (set $\lambda = 1$).

(i) Gaussian kernel $k(\mathbf{u}, \mathbf{v}) = \exp\left(-\frac{\|\mathbf{u}-\mathbf{v}\|^2}{2\sigma^2}\right)$ (set $\sigma = 1$)

(b) Submit your Python code.

Note that exceptionally for this problem you can use sklearn library.

2) **Principal Components Analysis (25 pts).**

In Principal Components Analysis (PCA), we project the data $X = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N | \mathbf{x}_i \in \mathbb{R}^D\}$ into K ($K < D$) orthogonal directions, which maximizes the following projection variance:

$$\max_{\substack{A \\ A^T A = I_k}} \sum_{k=1}^{K} \mathbf{a}_k^T S \mathbf{a}_k \tag{1}$$

where $S = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T \in \mathbb{R}^{D \times D}$ is the data covariance matrix, transformation matrix $A = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_K \end{bmatrix} \in \mathbb{R}^{D \times K}$ and $\mathbf{a}_k^T \mathbf{x}_i$ is the projection of the i-th data point into the k-th direction. Suppose S has the eigenvalue decomposition $S = U \Lambda U^T$ where $U = \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_D \end{bmatrix} \in \mathbb{R}^{D \times D}$ and $U^T U = I_D$; diagonal matrix $\Lambda = diag(\lambda_1, \lambda_2, \cdots, \lambda_D)$ and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D$. Denote $\mathbf{w}_k = U^T \mathbf{a}_k$ and $W = \begin{bmatrix} \mathbf{w}_1 & \mathbf{w}_2 & \cdots & \mathbf{w}_K \end{bmatrix} \in \mathbb{R}^{D \times K}$, then we get the following optimization problem from (1):

$$\max_{\substack{W \\ W^T W = I_k}} \sum_{k=1}^{K} \mathbf{w}_k^T \Lambda \mathbf{w}_k \tag{2}$$

(a) (10 pts) We denote

$$h_j = \sum_{k=1}^{K} \|\mathbf{w}_k^{(j)}\|^2$$

i.e., the square of L_2 norm of the j-th row vector in W. Prove that $0 \leq h_j \leq 1$ and $\sum_{j=1}^{D} h_j = K$.

(b) (10pts) Prove that

$$\max_{\substack{W \\ W^T W = I_k}} \sum_{k=1}^{K} \mathbf{w}_k^T \Lambda \mathbf{w}_k = \max_{\substack{h_j \\ W^T W = I_k}} \sum_{j=1}^{D} h_j \lambda_j \tag{3}$$

(c) (5pts) What are the optimal h_j in (3)? Show that $\mathbf{a}_k = \mathbf{u}_k$ ($k = 1, \cdots, K$) is a solution of (3).