Bhishan Poudel

225/250

HW Assignment 7 (Due by 10:30am on Nov 30)

1 Theory (140 points)

50 1. [**Normalized Kernels**, 50 + 25 points]
Let $K : X \times X \to R$ be a kernel function defined over a sample space X.

(a) Prove that the function below (a *normalized kernel*) is a valid kernel.

$$\frac{K(x,y)}{\sqrt{K(x,x)K(y,y)}}$$

(b) Why it would not make sense to normalize a Gaussian kernel?

(c) [**Bonus**] Which types of input data would benefit from normalizing the kernel function? Explain why, and provide real world examples.

50 2. [**Support Vector Machines**, 50 points]
Prove that the sum of slacks $\sum \xi_n$ from the objective function of the SVM formulation with soft margin is an upper bound on the number of misclassified training examples.

20 3. [**Max Margin Hyperplanes**, 20 points]
Consider the constrained optimization SVM problem for the separable case shown on slide 12. Show that, if the 1 on the right-hand side of the inequality constraint is replaced by some arbitrary constant $\gamma > 0$, the resulting maximum margin hyperplane is unchanged.

40 4. [**Kernel Techniques**, 20 + 20 points]
(a) Show that if $k_1(\mathbf{x}, \mathbf{y})$ and $k_2(\mathbf{x}, \mathbf{y})$ are valid kernel functions, then $k(\mathbf{x}, \mathbf{y}) = k_1(\mathbf{x}, \mathbf{y})k_2(\mathbf{x}, \mathbf{y})$ is also a valid kernel.
(b) **(*)** Show that if A is a symmetric positive semidefinite matrix, then $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T A\mathbf{y}$ is a valid kernel. *Hint: Inderjit Dhillon's Linear Algebra Background describes some useful properties of symmetric positive semidefinite matrices.*

20 5. [**Positive Definite Matrices (*)**, 20 points]
Show that a diagonal matrix $\mathbf{W}$ whose elements satisfy $0 < W_{ii} < 1$ is positive definite. Show that the sum of two positive definite matrices is itself positive definite.

50 30 6. [**Large Margin Perceptron (*)**, 30 points]
Let $\mathbf{u}$ be a current current vector of parameters and $\mathbf{x}$ and $\mathbf{y}$ two training examples such that $\mathbf{u}^T(\mathbf{x} - \mathbf{y}) < 1$. Use the technique of Lagrange multipliers to find a new vector of parameters $\mathbf{w}$ as the solution to the convex optimization problem below:

minimize:
$$J(\mathbf{w}) = \tfrac{1}{2}\|\mathbf{w} - \mathbf{u}\|^2$$

subject to:
$$\mathbf{w}^T(\mathbf{x} - \mathbf{y}) \geq 1$$

2 Text Classification (100 points)

Train and test the SVM algorithm on the *Spam vs. Non-spam* and *Atheism vs. Religion* classification problems, using the datasets provided for the previous assignment. Use a linear kernel, with the cost parameter $C = 5$. Report and compare the accuracy of the trained SVM models with the perceptron and average perceptron accuracies from the previous assignment.

3 Digit Recognition (200 points)

In this exercise, you are asked to run an experimental evaluation of SVMs and the perceptron algorithm, with and without kernels, on the problem of classifying images representing digits.

1. The UCI Machine Learning Repository at *www.ics.uci.edu/~mlearn* maintains datasets for a wide variety of machine learning problems. For this assignment, you are supposed to work with the Optical Recognition of Handwritten Digits Data Set. The webpage for this dataset is at:
   http://archive.ics.uci.edu/ml/datasets/Optical+Recognition+of+Handwritten+Digits
   The actual dataset is located at:
   http://archive.ics.uci.edu/ml/machine-learning-databases/optdigits/

   Read the description of the dataset. Download the training set `optdigits.tra` and the test set `optdigits.tes`. Use the first 1000 examples in `optdigits.tra` for *development* and the rest of 2823 examples for *training*. Use all 1797 examples in `optdigits.tes` for *testing*. Scale all the features between [0, 1], as discussed in class, using the min and max computed over the training examples. Create training files for each of the 10 digits, setting the class to 1 for instances of that digit, and to -1 for instances of other digits, i.e. *one-vs-rest* scenario.

2. Train first the linear perceptron, with the number of epochs set to $T \in \{1, 2, ..., 20\}$. After training each linear perceptron, normalize the learned weight vector. Select for T the value that obtains the best overall accuracy on the development data, and use this value for the remaining perceptron experiments.

   Run experiments with the linear and kernel perceptron algorithms. For the kernel perceptron, experiment with polynomial kernels $k(\mathbf{x}, \mathbf{y}) = (1 + \mathbf{x}^T\mathbf{y})^d$ with degrees $d \in \{2, 3, 4, 5, 6\}$, and with Gaussian kernels $k(\mathbf{x}, \mathbf{y}) = exp(-||\mathbf{x} - \mathbf{y}||^2/2\sigma^2)$ with the width $\sigma \in \{0.1, 0.5, 2, 5, 10\}$. For each hyper-parameter value, you will have trained 10 models, one for each digit. In order to compute the label for a test or development example, you will run the 10 trained models and output the label that obtains the highest score. Compute the accuracy on the development data and identify the hyper-parameter value that obtains the best accuracy. Use the tuned hyper-parameter (d for poly-kernel, σ for Gaussian) to compute the overall performance on the test data.

   For each of the three perceptrons (linear, poly kernel and Gaussian kernel) report the total training time, the overall accuracy, and the number of support vectors. Show and compare the corresponding 4 confusion matrices. Which digit seems to be the hardest to classify? Which perceptron / kernel combination achieves the best performance? Which algorithms are slower at training time, and why?

3. Run the same experiments using SVMs instead of perceptrons, i.e. linear SVMs and SVMs with polynomial and Gaussian kernels. Use the same tuning scenarios for the hyper-parameters of the polynomial and Gaussian kernels. Use $C = 1$ in all SVM experiments. Report the same types of results and analysis as above, and compare with the perceptron results.

4 Tools

You are free to use MATLAB, R, or packages written in C++/Java/Python such as SVM-LIGHT (C), LIBSVM (C++, Java), or SCIKIT-LEARN (Python) to complete the implementation part of this assignment. Their web sites contain plenty of documentation on how to use them. If you use SCIKIT-LEARN, the following functionality from the `sklearn.svm` will be useful:

1. `SVC()`: This is the main class used for SVM classification models. Its implementation is based on LIBSVM. Make sure that you properly map the SVM hyper-parameters to the parameters in the constructor of this class. For example, the *gamma* parameter in the constructor corresponds to our $1/2\sigma^2$ coefficient in the Gaussian kernel. The formulas for the kernels implemented by SVC are described in this User Guide.

2. `decision_function(x)`: Once the classifier is trained, this will compute the distance between a sample `x` and the decision hyperplance. This is the quantity that you can use to determine the highest scoring class when training the 10 $one - vs - rest$ classifiers: once a classifier is trained for all 10 digits, given a sample `x` you compute this quantity for all 10 classifiers and select the class that corresponds to the classifier with largest decision function value.

3. `fit()`: This is the function used to train the classifier.

4. `predict(x)`: This is used to calculate the (binary) label for sample `x`.

LIBSVM, and therefore SCIKIT-LEARN too, already implement the *one-vs-rest* classification scheme. In this scheme, you can directly use the training dataset with the 10 original labels, and SCIKIT-LEARN will train the 10 binary classifiers for you. You can use this capability for this assignment, however bonus points will be given if you train the 10 binary classifiers directly, as described for the perceptron algorithm above, by creating a binary training dataset for each class.

5 Submission

Turn in a hard copy of your homework report at the beginning of class on the due date. Electronically submit on Blackboard a hw07.zip file that contains the hw07 folder in which you place the code and the datasets. Make sure you include a README.txt file explaining how the code is supposed to be used to replicate the results included in the report. The screen output produced when running the code should be redirected to (saved into) an **output.txt** file.

On a Linux system, creating the archive can be done using the command:

```
> zip -r hw07.zip hw07
```

Please observe the following when handing in homework:

1. Structure, indent, and format your code well.
2. Use adequate comments, both block and in-line to document your code.
3. **Do not submit third-party ML packages on Blackboard!** Just explain in the REAMDE file how you use external packages.
4. Make sure your code runs correctly when used in the directory structure shown above.
5. **Type and nicely format the project report**, including discussion points, tables, graphs etc. so that it is presentable and easy to read.
6. Working code and/or correct answers is only one part of the assignment. The project report, including discussion of the specific issues which the assignment asks about, is also a very important part of the assignment. Take the time and space to make an adequate and clear project report. On the non-programming learning-theory assignment, clear and complete explanations and proofs of your results are as important as getting the right answer.

HW 07

Bhishan Poudel

65/50

A1a

Normalized kernel is a valid kernel.

5/5

Let $K(x,y)$ is a valid kernel, then, there exist a feature map $\phi(\cdot)$ in some Hilbert space that,

$$K(x,y) = \phi(x) \cdot \phi(y)$$

Now, $$\hat{K}(x,y) = \frac{K(x,y)}{\sqrt{K(x,x) \cdot K(y,y)}}$$

$$= \frac{\phi(x) \cdot \phi(y)}{\|\phi(x)\| \, \|\phi(y)\|} = \hat{\phi}(x) \cdot \hat{\phi}(y)$$

where $\hat{\phi}(x) = \frac{\phi(x)}{\|\phi(x)\|}$

since $K(x,y)$ is kernel $\Rightarrow \phi(x) \cdot \phi(y) \geq 0$

also, denominator $\Rightarrow \|\phi(x)\| \, \|\phi(y)\| \geq 0$

$\therefore \hat{K}(x,y)$ is non-negative.

also $\hat{K}(x,y)$ is symmetric.

$\therefore$ $\hat{K}(x,y)$ is a valid kernel.

25/25

Q1N1b Gaussian function,

Gaussian fn

$$g(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\|x-\mu\|^2}{2\sigma^2}}$$

(normalized gaussian function when $y=\mu$)

the gaussian kernel is,

gaussian kernel

$$K(x,y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}} = e^{-\gamma\|x-y\|^2}$$

where $\gamma = \frac{1}{2\sigma^2}$

$$K(x,x) = e^{-0} = 1$$

$K(x,x) = 1$ for gaussian kernel.

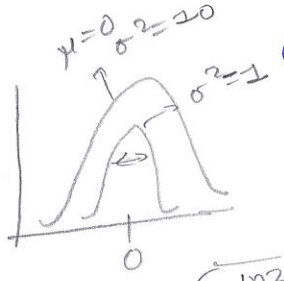
Since gaussian kernel is already normalized ($K(x,x)=1$), it would make no sense to normalize gaussian kernel.

10 some of the kernels used in practice are

linear kernel $K(x,y) = x^T y + c = \langle x, y \rangle + c = \vec{x} \cdot \vec{y} + c$

polynomial kernel $K(x,y) = (\alpha x^T y + c)^d$ α = slope, d = degree, c = const

Gaussian kernel $K(x,y) = e^{-\frac{1}{2\sigma^2}\|x-y\|^2} = e^{-\gamma\|x-y\|^2}$



(σ or γ determines the width of the gaussian curve.

overestimate σ → behave like linear kernel

underestimate σ → lack regularization and high sensitive to noise

tuning of σ → very important, otherwise very low performance

exponential kernel $K(x,y) = e^{-\frac{1}{\sigma}\|x-y\|}$

sigmoid kernel $K(x,y) = \tanh(\alpha x^T y + c)$

Gaussian kernels are already normalized and we need to normalize polynomial kernels.

cosine Normalization →
- Reduces dimension of data by 1 (since it projects all the data inside a unit radius sphere)
- → Good for high dimensional input data e.g. optdigits with 64 features
- → Not good for low dimensional data e.g. Iris data with only ~~two~~ four features.

ON2

prove: sum of slacks $\leq \xi_n$ is an upper bound on the number of misclassified training examples.

solution:

Input data X

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_i \\ \vdots \\ x_N \end{bmatrix}$$

Target t

$$\begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_i \\ \vdots \\ t_N \end{bmatrix} \qquad t \in \{-1, 1\}$$

$x_i \in \mathbb{R}^d$

$x_i \in \{x_{i1}, x_{i2}, \cdots, x_{id}\}$ d = dimension of features.

hypothesis $h_i = w^T x_i + b$

$\forall i \in \{1, \cdots N\}$

algorithm : constraint $t_i (w^T x_i + b) \geq 1 - \xi_i$

$$J(w, b) = \frac{1}{2} \|w\|^2$$

minimization objective

minimize $J(w, b) = \frac{1}{2} \|w\|^2$

with constraint $t_n (w^T x_n + b) \geq 1 - \xi_n$

consider $t_i = +1$

(correct classification)

$w^T x_i + b \geq 0$

$t_i (w^T x_i + b) \geq 0 \quad \because t_i = +1$

$t_i (w^T x_i + b) \geq 1 - \xi_i \; ; \; \xi_i \geq 0$

$\xi_i \geq 1$

$+ve \geq 1 - \xi_i$

since constraint want $\xi_i \geq 0$

ξ_i cannot be greater than 1

$0 \leq \xi_i \leq 1$

$t_i (w^T x_i + b) \geq 1$ if $\xi_i = 0$

(hard margin)

aside:

for perceptron

$h_i = sgn(w^T x_i + b)$

$h_i = t_i (w^T x_i + b)$

and if

$h_i \neq t_i$ we

update w, b

consider $t_i = +1$

(misclassification)

$w^T x_i + b < 0$

$t_i (w^T x_i + b) < 0 \quad \because t_i = +1$

constraint:

$t_i (w^T x_i + b) \geq 1 - \xi_i$

$-ve \geq 1 - \xi_i$

$\xi_i \geq 1 + (+ve)$

$\xi_i \geq 1$

consider $t_i = -1$ (correct classification)

$w^T x_i + b \leq 0$

$t_i (w^T x_i + b) \geq 0 \quad \because t_i = -1$

constraint:

$t_i (w^T x_i + b) \geq 1 - \xi_i$

$+ve \geq 1 - \xi_i$

$\xi_i \geq 1 - (+ve)$

$0 \leq \xi_i \leq 1$

and $t_i (w^T x_i + b) \geq 1$ if $\xi_i = 0$

$\Rightarrow$ for correct classification

$0 \leq \xi_i \leq 1$

$t_i = -1$ (misclassified)

$w^T x_i + b \geq 0$

$t_i (w^T x_i + b) < 0 \quad \because t_i = -1$

constraint:

$t_i (w^T x_i + b) \geq 1 - \xi_i$

$-ve \geq 1 - \xi_i$

$\xi_i \geq 1 + (+ve)$

$\xi_i \geq 1$

$\Rightarrow$ for misclassification

$\xi_i \geq 1$

now, for correct classification

$\xi_i \geq 0$

$\sum_{i \in \text{correct}} \xi_i \geq 0$

for misclassified

$\xi_i > 1$

$\sum_{i \in \text{misclassified}} \xi_i > \sum_{i \in \text{misclassified}} 1$

$\sum_{\text{misclassified}} \xi_i > \#\text{ of misclassified}$

$\therefore \sum_{i \in \text{correct}} \xi_i + \sum_{i \in \text{misclassified}} \xi_i > \#\text{ of misclassified}$

$\therefore \boxed{\sum_{i=1}^{N} \xi_i > \#\text{ of misclassified}}$

Q.E.D.

This shows that sum of slacks (ξ_i) over all the N training examples is greater than total number of misclassified examples, hence, sum of slacks is an upper bound to the number of misclassified examples.

Aside: The constrained optimization of SVM algorithms $t_i(w^T x_i + b) \geq 1$ is called Hard margin SVM, and, $t_i(w^T x_i + b) \geq 1 - \xi_i$ is called Soft margin SVM. They both are non-trivial algorithm (unlike vanilla (perceptron) and use Lagrange multipliers to solve the optimization.

Q3 max margin hyperplane

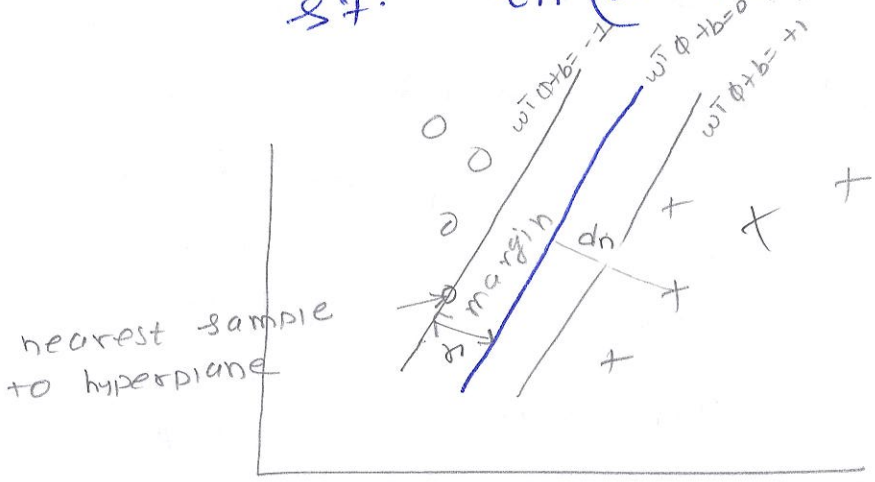
20/20

For the linearly separable dataset, the constrained optimizer for SVM is given by

$$\min \; J(w,b) = \frac{1}{2}\|w\|^2$$

(hard margin SVM)

$$\text{s.t.} \quad t_n\left(w^T \phi(x_n) + b\right) \geq 1 \quad \forall n \in \{1, \dots, N\}$$



distance to any sample,

$$d_n = \frac{t_n\left(w^T \phi(x_n) + b\right)}{\|w\|}$$

from slides, (lecture 07, page 5)

$$\text{margin} = \min_n \frac{t_n\, w^T \phi(x_n) + b}{\|w\|}$$

we find parameters w and b that maximizes margin,

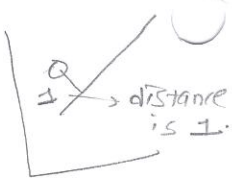
$$w^*, b^* = \underset{w,b}{\text{argmax}} \; \frac{1}{\|w\|} \cdot \min_n \; t_n\left(w^T \phi(x_n) + b\right)$$

Since w^*, b^* are obtained from argmax, they do not depend on rescaling of w, b.

So, without loss of generality, we can say that the nearest distance to the sample from hyperplane is 1.

i.e. for closest point(s),

$$t_n(w^T \phi(x_n) + b) = 1$$



then, our optimizing constraint becomes,

$$t_n(w^T \phi(x_n) + b) \geq 1 \quad \forall n \{1, \cdots, N\}$$

[Aside: previously we used

$t_n(w^T \phi + b) = 1$

$t_n h_n = 1$]

(for perceptron)

$w^T \phi(x_n) + b > 0$ if this +1

$w^T \phi(x_n) + b < 0$ if this −1

$h_n = \text{sgn}(w^T \phi(x_n) + b)$

now, with inclusion of this constraint, the new optimization problem for SVM becomes,

minimize

$$J(w, b) = \frac{1}{2} \|w\|^2$$

$$\text{s.t.} \quad t_n(w^T \phi(x_n) + b) \geq 1 \quad \forall n \in \{1, \cdots N\}$$

here we have use the margin $\gamma = 1$.

we have to show that if $\gamma \neq 0$, then also the decision hyperplane is unchanged.

Solution :

the distance to the nth sample from the decision hyperplane is,

$$d_n = \frac{t_n(w^T \phi(x_n) + b)}{\|w\|} \quad \} \text{ —— (1)}$$

if we rescale the weight vector to 1, $\|w\| = 1$, then,

$$d_n = t_n(w^T \phi(x_n) + b) \quad \} \text{ —— (2)}$$

then, maximizing margin algorithm becomes,

$$\max_{w,b} \min_n d_n$$

$$\text{s.t. } \|w\| = 1 \quad \} \text{ —— (3)}$$

let γ be the maximum margin to closest training example,

$$\forall_n \quad t_n(w^T \phi(x_n) + b) \geq \gamma$$

$$\Rightarrow \quad t_n\left(\frac{w^T}{\gamma}\phi + \frac{b}{\gamma}\right) \geq 1$$

$$\Rightarrow \quad t_n(V^T \phi + B) \geq 1 \quad \} \text{ —— (4)}$$

where $V = \frac{w}{\gamma}$, $B = \frac{b}{\gamma}$, γ = margin to the closest example.

then, $V = \frac{w}{\gamma}$

$$\|V\|^2 = \left\|\frac{w}{\gamma}\right\|^2 = \frac{1}{\gamma^2}\|w\|^2$$

$$\boxed{\|V\|^2 = \frac{1}{\gamma^2}} \quad \text{since } \|w\| = 1$$

Hence, our original optimization problem,

$$\max_{w,b} \min_n d_n = t_n(w^T \phi(x_n) + b)$$

reduces to maximizing $\frac{1}{\gamma^2}$

i.e. minizing $\|V\|^2$

i.e. minizing $\frac{1}{2}\|V\|^2$

which we can write as,

$$\min \frac{1}{2}\|V\|^2$$

$$s.t. \quad t_n(V^T \phi(x_n) + B) \geq 1. \quad \text{(from 4)}$$

which is same as,

$$\min J(w,b) = \frac{1}{2}\|w\|^2$$

$$s.t. \quad t_n(w^T \phi(x_n) + b) \geq 1 \quad \forall n \in \{1, N\}$$

Q.E.D.

q1 4a — 30/40

product of two kernels is a kernel.

Let ϕ_1, ϕ_2 are the feature map of kernels K_1, K_2 then,

20/20

$$K_1(x_1,x_2)\, K_2(x_1,x_2) = (\phi_1(x_1)\ \phi_1(x_2))\ (\phi_2(x_1)\ \phi_2(x_2))$$

$$= \left(\sum_{i=1}^{\infty} f_i(x_1)\, f_i(x_2)\right)\left(\sum_{j=1}^{\infty} g_j(x_1)\, g_j(x_2)\right)$$

$$= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} f_i(x_1)\, f_i(x_2)\ g_j(x_1)\ g_j(x_2)$$

$$= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} (f_i(x_1)\, g_j(x_1))\ (f_i(x_2)\, g_j(x_2))$$

$$= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} h_{i,j}(x_1)\ h_{i,j}(x_2)$$

where $h_{i,j}(x)$ is a feature vector in feature space ϕ_3 such that

$$h_{i,j}(x) = f_i(x)\, g_j(x)$$

so, $K_1(x_1,x_2) \cdot K_2(x_2,x_3) = \phi_3(x_1)^T\ \phi_3(x_2)$

is a valid kernel.

~~Q4b~~ $x^T A y$ is valid kernel if A is sym PSD matrix.

Soln: Given $A \rightarrow$ symmetric positive semi-define

$A^T A$ is also sym PSD (from internet Shillan's reference)

Now,

Let, $K(x,y) = x^T A^T A y$

$= (Ax)^T (Ay)$ $\because (AB)^T = B^T A^T$

$= \phi(x)^T \phi(y)$

where we define feature map

$\phi(3) = A3$

$\therefore K(x,y)$ is a valid kernel.

Q.NO 5.a

20/20

Diagonal matrix with positive elements is posisitve definite matrix.

soln: Let $W = W_{n,n} \in \mathbb{R}^{n \times n}$ is a diagonal matrix then.

$$W = \begin{bmatrix} w_{11} & & 0 \\ & w_{22} & \\ 0 & & \ddots \; w_{nn} \end{bmatrix}_{n,n}$$

Let $\vec{x}$ be a n-dimensional column vector,

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}_{n,1}$$

then, the inner product of diagonal matrix W associated with $\vec{x}$ is,

$$x^T W x = \underset{1,n}{[x_1 \; x_2 \cdots x_n]} \begin{bmatrix} w_{11} & & 0 \\ & \ddots & \\ 0 & & w_{nn} \end{bmatrix}_{n,n} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}_{n,1}$$

$$= \sum_{j=1}^{n} w_j x_j^2$$

where $w_j = w_{ii}$ (diagonal elements)

$$x^T W x = \sum_{j=1}^{n} w_j x_j^2 \quad = +ve \text{ number}$$

here all the diagonal elements w_j are positive ($0 < w_{ii} < \infty \quad \forall\, i \in (1, n)$).

then the RHS quantity is positive

and hence W is a positive define matrix.

Alliter: $W = \text{diag}(w_{11}, w_{22}, \ldots, w_{nn})$, $w_{ii} > 0$

let $\sqrt{W} = \text{diag}(\sqrt{w_{11}}, \sqrt{w_{22}}, \ldots, \sqrt{w_{nn}})$

$$x^T W x = (x^T \sqrt{W}^T)(\sqrt{W} x)$$
$$= (\sqrt{W} x)^T (\sqrt{W} x)$$
$$= \| \sqrt{W} x \|^2 \geq 0$$

$\therefore$ W is PSD.

Ans 6 Sum of two PSD matrices is PSD matrix.

Let W is a positive semi-definite matrix

$\Rightarrow x^T W x \geq 0 \quad \forall x \in \mathbb{R}^n$

Let V is a positive semi-def matrix

$\Rightarrow x^T V x \geq 0$

now, $x^T W x + x^T V x \geq 0$

or, $x^T (W+V) x \geq 0$ (distributive property of matrix)

this means that sum of two matrices $W+V$ is also positive semi-definite.

proved

Q1.6 $w = u + \alpha(x-y)$

then, $J(w) = \frac{1}{2}\|w-u\|^2$

$= \frac{1}{2}\|\alpha(x-y)\|^2$

constraint: $w^T(x-y) = u^T(x-y) + \alpha\|x-y\|^2 \geq 1$

for, $\alpha\|x-y\|^2 \geq 1 - u^T(x-y)$

$\alpha \geq \frac{1-u^T(x-y)}{\|x-y\|^2}$

Then, optimization problem:

$\min_{\alpha} J(\alpha) = \frac{1}{2}\alpha^2\|x-y\|^2$

s.t. $\alpha \geq \frac{1-u^T(x-y)}{\|x-y\|^2}$

Again, $J(\alpha)$ will be minimum when α will be minimum.

But $\min \alpha = \frac{1-u^T(x-y)}{\|x-y\|^2}$

so, $\alpha = \frac{1-u^T(x-y)}{\|x-y\|^2}$

then, the w is given by (eliminate α)

$w = u + \alpha(x-y)$

$$w = u + \frac{1-u^T(x-y)}{\|x-y\|^2}(x-y)$$

Ans

Q6 Lagrange multipliers

Here, the optimization problem is,

$$\underset{w}{\text{minimize}} \quad J(w) = \frac{1}{2}\|w-u\|^2$$

$$\text{s.t.} \quad w^T(x-y) \geq 1$$

$$wx - wy \geq 1$$

$$1 - wx + wy \leq 0$$

primal Lagrangian

$$L_p(w, x, y, \alpha) = \frac{1}{2}(w-u)^2 + \alpha(1 - wx + wy)$$

$\alpha \geq 0$ is Lagrange multiplier

Set gradients to zero

$$0 = \frac{\partial L_p}{\partial w} = w - u - \alpha x + \alpha y \Rightarrow w = u + \alpha x - \alpha y = u + \alpha(x-y)$$

$$0 = \frac{\partial L_p}{\partial \alpha} = 1 - wx + wy \Rightarrow 1 = w(x-y) \Rightarrow x - y = \frac{1}{w}$$

then, $w = u + \frac{\alpha}{w} \Rightarrow w^2 = uw + \alpha \Rightarrow w^2 - u \cdot w - \alpha = 0$

$$w^* = \frac{u \pm \sqrt{u^2 + 4\alpha}}{2}$$

This is NOT solution look LHS

In gradient descent algorithm we initialize $\vec{u} = 0$ and $\alpha \geq 0$ and find the optimum value of w^*.