SI151: Optimization and Machine Learning
Reference Solutions of Final Exam

June 10, 2021

I Basics [20 points]

Note: in the following questions, you may mark one or more than one of the choices.

1. [**2 points**] Linear regression estimator has the smallest variance among all unbiased estimators.

   (a) True

   (b) False

Solution
B

2. [**2 points**] Since classification is a special case of regression, logistic regression is a special case of linear regression.

   (a) True

   (b) False

Solution
B

3. [**2 points**] The training error of 1-nearest neighbor classifier is 0.

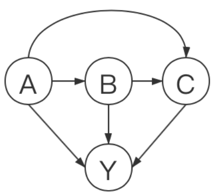
   (a) True

   (b) False

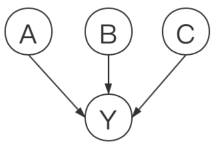
Solution
A

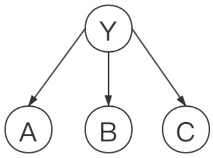
4. [**2 points**] Suppose that you have a dataset with 3 categorical input attributes A, B and C. There is one categorical output attribute Y. You are trying to learn a Naive Bayes Classifier for predicting Y. Which of these Bayes Net diagrams represent(s) the naive bayes classifier assumption?

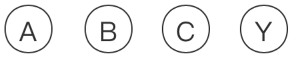
   (a)



   (b)



(c) 

(d) 

Solution
C

5. [**2 points**] In each round of AdaBoost, the misclassification penalty for a particular training observation is increased going from round t to round $t+1$ if the observation was:

   (a) classified incorrectly by the weak learner trained in round t.

   (b) classified incorrectly by the full ensemble trained up to round t.

   (c) classified incorrectly by a majority of the weak learners trained up to round t.

Solution
A

6. [**2 points**] AdaBoost minimizes an exponential loss function.

   (a) True

   (b) False

Solution
A

7. [**2 points**] What statement(s) are true about the expectation-maximization (EM) algorithm?

   (a) It requires some assumption about the latent probability distribution.

   (b) Comparing to a gradient descent algorithm that optimizes the same objective function as EM, EM may only find a local optima whereas the gradient descent will always find the global optima.

   (c) The EM algorithm maximizes a lower bound of the marginal likelihood $P(\mathcal{D}; \boldsymbol{\theta})$

   (d) The algorithm assumes that some of the data generated by the probability distribution is not observed.

Solution
A, C, D

8. [**2 points**] The SVM learning algorithm is guaranteed to find the globally optimal hypothesis with respect to its object function.

   (a) True

   (b) False

Solution
A

9. [**2 points**] Which statement(s) are true about the K-means algorithm?

   (a) It is a clustering algorithm.

   (b) It is an EM algorithm.

(c) It assumes the data is from a mixture of Gaussian distributions.

(d) It is a soft EM algorithm, where all possible hidden attributes are considered in the E step.

(e) It is guaranteed to converge to the global optimum.

(f) It is a convex optimization problem.

Solution
A, B, C

10. **[2 points]** Query strategy plays a key role in active learning. Generally, the following query strategies can be selected: uncertainty sampling, query-by-committee, expected model change, expected error reduction, variance reduction, density-weighted methods. Which of the following option(s) is(are) reasonable method(s) of query strategies?

(a) Least confident method, which is to select samples that have a low maximum classification probability.

(b) Margin sampling method, which is to select samples of data that can easily be classified into two categories, or that have a similar probability of being classified into two categories.

(c) Entropy method, which is to select samples of data that have high entropy in a particular system. (The definition of entropy is $-\sum_i P_\theta(y_i \mid x) \cdot \ln P_\theta(y_i \mid x)$.)

(d) Expected loss method, which is to select samples of data that will cause the loss function to reduce the least by adding a sample.

Solution
A, B, C

II Regression and Probability Estimation [12 points]

Consider real-valued variables X and Y, in which Y is generated conditional on X according to

$$Y = aX + \epsilon, \text{ where } \epsilon \sim \mathcal{N}(0, \sigma^2).$$

Here ϵ is an independent variable, called a noise term, which is drawn from a Gaussian distribution with mean 0, and variance σ^2. This is a single variable linear regression model, where a is the only weight parameter. The conditional probability of Y has distribution $p(Y|X, a) \sim \mathcal{N}(aX, \sigma^2)$, so it can be written as:

$$p(Y|X, a) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(Y - aX)^2\right).$$

The following questions are all about this model.

1. **[4 points]** Assume we have a training dataset of n pairs (X_i, Y_i), $i = 1, 2, ..., n$. Which one(s) of the following equations correctly represent(s) the Maximum Likelihood Estimation (MLE) problem for estimating a? (You may mark one or more than one of the choices.)

   (a) $\arg\max_a \sum_i \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(Y_i - aX_i)^2\right)$

   (b) $\arg\max_a \prod_i \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(Y_i - aX_i)^2\right)$

   (c) $\arg\max_a \sum_i \exp\left(-\frac{1}{2\sigma^2}(Y_i - aX_i)^2\right)$

   (d) $\arg\max_a \prod_i \exp\left(-\frac{1}{2\sigma^2}(Y_i - aX_i)^2\right)$

   (e) $\arg\max_a \frac{1}{2}\sum_i (Y_i - aX_i)^2$

   (f) $\arg\min_a \frac{1}{2}\sum_i (Y_i - aX_i)^2$

   Solution

   B, D, F

2. **[4 points]** Derive the maximum likelihood estimate of the parameter a in terms of the training data (X_i, Y_i), $i = 1, 2, ..., n$. You are recommended to start with the simplest form of the problem you found above.

   Solution

$$\begin{aligned}
0 &= \frac{\partial}{\partial a}\left[\frac{1}{2}\sum_i (Y_i - aX_i)^2\right] && (1)\\
&= \sum_i (Y_i - aX_i)(-X_i) && (2)\\
&= \sum_i aX_i^2 - X_iY_i && (3)\\
a &= \frac{\sum_i X_iY_i}{\sum_i X_i^2} && (4)
\end{aligned}$$

3. **[4 points]** Let's put a prior on a, for example, $a \sim \mathcal{N}(0, \lambda^2)$, i.e.,

$$p(a|\lambda) = \frac{1}{\sqrt{2\pi}\lambda} \exp(-\frac{1}{2\lambda^2}a^2).$$

   (a) Under which case(s) that the estimated value with MLE and Maximum A Posterior (MAP) will become closer, in other words, $|a^{MLE} - a^{MAP}|$ will decrease? (You may mark one or more than one of the choices.)

      i. As $\lambda \to \infty$

      ii. As $\lambda \to 0$

      iii. Fix λ and as number of training samples $n \to \infty$

Solution
A, C

(b) Assume $\sigma = 1$, and a fixed prior parameter λ. Solve for the MAP estimate of a:

$$\arg\max_a \left[\log p(Y_1, ..., Y_n | X_1, ..., X_n, a) + \log p(a|\lambda)\right].$$

Your solution should be in terms of X_i's Y_i's and λ.

Solution

$$\begin{aligned}
\frac{\partial}{\partial a}[\log p(Y \mid X, a) + \log p(a \mid \lambda)] &= \frac{\partial \ell}{\partial a} + \frac{\partial \log p(a \mid \lambda)}{\partial a} && (5)\\
\frac{\partial \ell}{\partial a} &= -\sum_i (Y_i - aX_i)(-X_i) && (6)\\
&= \sum_i (Y_i - aX_i) X_i && (7)\\
&= \sum_i X_i Y_i - aX_i^2 && (8)\\
\frac{\partial \log p(a \mid \lambda)}{\partial a} &= \frac{\partial}{\partial a}\left[-\log(\sqrt{2\pi}\lambda) - \frac{1}{2\lambda^2}a^2\right] && (9)\\
&= -\frac{a}{\lambda^2} && (10)\\
\Rightarrow 0 &= \frac{\partial \ell}{\partial a} + \frac{\partial \log p(a)}{\partial a} && (11)\\
\Rightarrow 0 &= \left(\sum_i X_i Y_i - aX_i^2\right) - \frac{a}{\lambda^2} && (12)\\
\Rightarrow a &= \frac{\sum_i X_i Y_i}{(\sum_i X_i^2) + 1/\lambda^2} && (13)
\end{aligned}$$

III Linear Classification [10 points]

Given the input continuous variable X and the output categorical variable Y, suppose that:

- We know $P(Y = k) = \pi_k$ exactly.
- $P(X = \mathbf{x} \mid Y = k)$ is multivariate normal distribution with density:

$$f_k(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mu_k)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_k)}, \quad \mathbf{x} \in \mathbb{R}^p,$$

where μ_k is the mean of the inputs for category k and $\mathbf{\Sigma}$ is the covariance matrix.

Answer the questions below:

1. [**3 points**] What is the Bayes classifier (maximize the probability of category k, given the input $\mathbf{x}$)?

Solution

$$\begin{aligned} P(Y = k \mid \mathbf{X} = \mathbf{x}) &= \frac{f_k(\mathbf{x})\pi_k}{P(\mathbf{X} = \mathbf{x})} \\ &= C\pi_k e^{-\frac{1}{2}(\mathbf{x}-\mu_k)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_k)}, \end{aligned}$$

where C denotes a constant irrelevant to k.

2. [**3 points**] Please derive the linear discriminant function $\delta_k(\mathbf{x})$, and explain how to predict the category of input $\mathbf{x}$.

Solution

$$\begin{aligned} \log P(Y = k \mid \mathbf{X} = \mathbf{x}) &= \log C + \log \pi_k - \frac{1}{2}(\mathbf{x}-\mu_k)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_k) \\ &= \log C + \log \pi_k - \frac{1}{2}\left[\mathbf{x}^T \mathbf{\Sigma}^{-1}\mathbf{x} + \mu_k^T \mathbf{\Sigma}^{-1}\mu_k\right] + \mathbf{x}^T \mathbf{\Sigma}^{-1}\mu_k \\ &= C' + \log \pi_k - \frac{1}{2}\mu_k^T \mathbf{\Sigma}^{-1}\mu_k + \mathbf{x}^T \mathbf{\Sigma}^{-1}\mu_k \end{aligned}$$

Thus,

$$\delta_k(\mathbf{x}) = \log \pi_k - \frac{1}{2}\mu_k^T \mathbf{\Sigma}^{-1}\mu_k + \mathbf{x}^T \mathbf{\Sigma}^{-1}\mu_k$$

At an input $\mathbf{x}$, we predict the category with the largest $\delta_k(\mathbf{x})$.

3. [**4 points**] Show what is the decision boundary between category k and l given the input $\mathbf{x}$. For some vectors $\mathbf{w}$ and scalar b, the decision boundary can be expressed as $\mathbf{w}^T\mathbf{x} + b = 0$. Find the entries of the vector $\mathbf{w}$ and the value of b in terms of class priors and parameters.

Solution

The decision boundary is

$$\delta_k(\mathbf{x}) = \delta_l(\mathbf{x})$$

$$\begin{aligned} &\log \pi_k - \frac{1}{2}\mu_k^T \mathbf{\Sigma}^{-1}\mu_k + \mathbf{x}^T \mathbf{\Sigma}^{-1}\mu_k = \log \pi_l - \frac{1}{2}\mu_l^T \mathbf{\Sigma}^{-1}\mu_l + \mathbf{x}^T \mathbf{\Sigma}^{-1}\mu_l \\ &\log \frac{\pi_k}{\pi_l} - \frac{1}{2}\mu_k^T \mathbf{\Sigma}^{-1}\mu_k + \frac{1}{2}\mu_l^T \mathbf{\Sigma}^{-1}\mu_l + \mathbf{x}^T \mathbf{\Sigma}^{-1}(\mu_k - \mu_l) = 0 \end{aligned}$$

Thus, $b = \log \frac{\pi_k}{\pi_l} - \frac{1}{2}\mu_k^T \mathbf{\Sigma}^{-1}\mu_k + \frac{1}{2}\mu_l^T \mathbf{\Sigma}^{-1}\mu_l$, $w_i = \left[\mathbf{\Sigma}^{-1}(\mu_k - \mu_l)\right]_i$

IV Graphical Model [10 points]

Consider the following Bayesian Network, in which all variables are boolean.

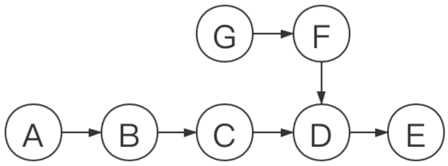


Figure 1: Bayesian network with seven boolean variables.

1. [**4 points**] Write the expression for the joint likelihood of the network in its factorized form.

> **Solution**
> $p(A, B, C, D, E, F, G) = p(A)p(B|A)p(C|B)p(D|C, F)p(E|D)p(F|G)p(G)$

2. [**3 points**] Let $X = \{C\}, Y = \{B, D\}, Z = \{A, E, F, G\}$. Is $X \perp Z|Y$?, If yes, explain why. If no, show a path from X to Z is not blocked.

> **Solution**
> No. The path $C \rightarrow D \rightarrow F$ is not blocked since D is head to head is observed.

3. [**3 points**] Directly prove that $A \perp C|B$ without using D-separation.

> **Solution**
> $p(A, C|B) = \frac{p(A,B,C)}{P(B)} = \frac{p(A,B)p(C|B)}{p(B)} = p(A|B)p(C|B)$

V Kernel Methods [8 points]

Kernel functions implicitly define some mapping function $\phi(\cdot)$ that transforms an input instance $x \in \mathbb{R}^d$ to a high dimensional feature space Q, by giving the form of dot product in $Q: K(x_i, x_j) = \phi(x_i)\dot{\phi}(x_j)$. Assume we use radial basis kernel function

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{1}{2}\|\mathbf{x}_i - \mathbf{x}_j\|^2\right).$$

1. **[4 points]** Prove that for arbitrary two input instances x_i and x_j, the squared Euclidean distance of their corresponding points in the feature space Q is less than 2, i.e.,

$$\|\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)\|^2 < 2.$$

Solution

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

2. **[2 points]** The dimensionality of the feature map generated by radial basis kernel is infinity.

   (a) True

   (b) False

Solution
A

3. **[2 points]** The dimensionality of the feature map generated by polynomial kernel (e.g., $K(x, y) = (1 + xy)^d$) is polynomial w.r.t. the power d of the polynomial kernel.

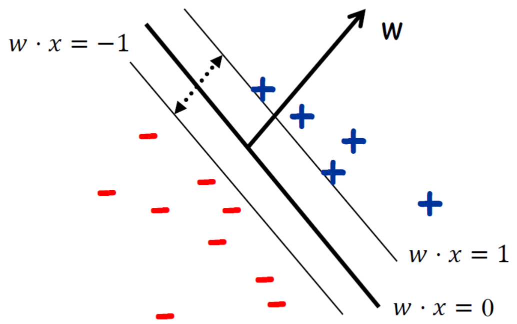
   (a) True

   (b) False

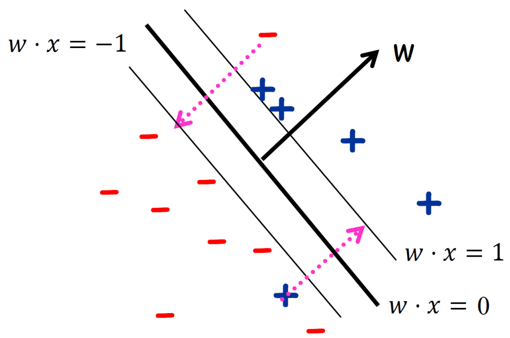
Solution
A

VI Support Vector Machines [10 points]

Support vector machines (SVM) are supervised learning models, that directly optimize for the maximum margin separator. Fig. 2 shows an example of maximum margin separator over a dataset $S = \{(x_i, y_i)\}_{i=1}^n$, in which $x_i \in \mathbb{R}^2$ and $y_i \in \{-1, 1\}$ denote the i-th sample and the i-th label ($\forall i$), respectively, in both separable case (Fig.2(a)) and non-separable case (Fig.2(b)). For simplicity, here we assume that the dataset S has been standardized, and thus the bias can be omitted in the linear model. In Fig. 2, "+" and "-" denote the samples with labels "1" and "-1", respectively, and $\mathbf{w}$ is the normal vector of the maximum margin separator $\mathbf{w}^\top x = 0$. You need to derive the linear optimization problem of SVM in both separable case and non-separable case.
Note: correctly giving the results without detailed derivation will get half the points.



(a) Separable case



(b) Non-separable case

Figure 2: Maximum margin separator.

1. **[3 points]** Derive the constraint optimization problem of SVM in the separable case shown in Fig. 2(a).

Solution
Let r be the margin between $\mathbf{w}^\top x = 0$ and $\mathbf{w}^\top x = 1$. Assume there are two points $x_0 \in \mathbb{R}^2$ and $x_1 \in \mathbb{R}^2$ on $\mathbf{w}^\top x = 0$ and $\mathbf{w}^\top x = 1$, respectively, and we make $x_1 - x_0$ paralleled with $\mathbf{w}$. Hence, we have the following equations:

$$\begin{cases} w^\top x_1 = 1, \\ w^\top x_0 = 0, \\ x_1 - x_0 = r \times \frac{\mathbf{w}}{||\mathbf{w}||_2}, \end{cases}$$

where $||\cdot||_2$ denotes the ℓ_2-norm. By multiplying $\mathbf{w}^\top$ on both sides of the third equation, and plugging the first two equations into it, we have

$$\begin{aligned} \mathbf{w}^\top (x_1 - x_0) &= r \times \frac{\mathbf{w}^\top \mathbf{w}}{||\mathbf{w}||_2} \\ 1 &= r \times ||\mathbf{w}||_2, \\ \Rightarrow \quad r &= \frac{1}{||\mathbf{w}||_2}. \end{aligned}$$

In the separable case, a maximum margin separator should satisfy the following three conditions:

- maximize the margin $r = \frac{1}{||\mathbf{w}||_2}$ over a dataset;
- put positive samples ($y_i = 1$) on one side of the separator, i.e., $\mathbf{w}^\top x_i \geq 1$;
- put negative samples ($y_i = -1$) on another side of the separator, i.e., $\mathbf{w}^\top x_i \leq -1$.

Therefore, the constraint optimization problem of SVM is

$$\begin{aligned} &\min_{\mathbf{w}} \ ||\mathbf{w}||_2^2, \\ &\text{s.t. } y_i \mathbf{w}^\top x_i \geq 1, \ \forall i \in \{1, 2, ..., n\}. \end{aligned}$$

2. **[3 points]** Extend the results in (a) to handle the non-separable case shown in Fig. 2(b).

Solution

To handle the non-separable case shown in Fig. 2(b), we need introduce the slack variable $\xi_i \geq 0$ $(\forall i)$ to move the possible outlier x_i to the correct side of the separator, and penalize the total amount of ξ_i in the objective function.

Thus, the optimization problem becomes

$$\begin{aligned} &\min_{\mathbf{w},\xi} ||\mathbf{w}||_2^2 + C\sum_{i=1}^{n}\xi_i, \\ &\text{s.t. } y_i\mathbf{w}^\top x_i \geq 1-\xi_i, \ \forall i, \\ &\quad \xi_i \geq 0, \forall i, \end{aligned}$$

where $C > 0$ is the regularization parameter.

3. [**4 points**] Show the unconstraint form of the above problem and determine the convexity. You need to explain the reason for your answer.

Solution

The unconstraint form of the above problem:

$$\min_{\mathbf{w}} ||\mathbf{w}||_2^2 + C\sum_{i=1}^{n}(1-y_i\mathbf{w}^\top xi_i)_+,$$

where $(\cdot)_+ = \max(0,\cdot)$.

This is a convex optimization problem due to the following three reasons:

- $||\mathbf{w}||_2^2$ is convex because its Hessian matrix is $\mathbf{I} \succeq 0$;
- $(1-y_i\mathbf{w}^\top xi_i)_+$ is convex because is the pointwise maximum of two affine functions;
- The objective is convex because it is a summation of many convex functions.

VII Principal Component Analysis [9 points]

Given 3 data points in 2D space: $(1,1)$, $(2,2)$, $(3,3)$, answer the following questions:

1. [**3 points**] What is the first principle component?

> **Solution**
> $(1/\sqrt{2}, 1/\sqrt{2})^\top$ (the negation is also correct)

2. [**3 points**] If we want to project the original data points into 1D space by principle component you choose, what is the variance of the projected data?

> **Solution**
> $4/3 = 1.33$

3. [**3 points**] For the projected data in 1D space, now if we represent them in the original 2D space, what is the reconstruction error?

> **Solution**
> 0

VIII Neural Networks [9 points]

Consider the network shown in the figure. All of the hidden units use the rectified linear unit (ReLU): $h_i = \max(z_i, 0)$. We are trying to minimize a cost function C which depends only on the activation of the output unit y. The unit h_1 (marked with $\star$) receives an input of -1 on a particular training case, so its output is 0. Based only on this information, which of the following weight derivatives are guaranteed to be 0 for this training case? Write TRUE or FALSE for each. (Hint: don't work through the backpropagation computations, instead, think about what do the partial derivatives really mean.)
Note: correct answers without explanation will get half the points.

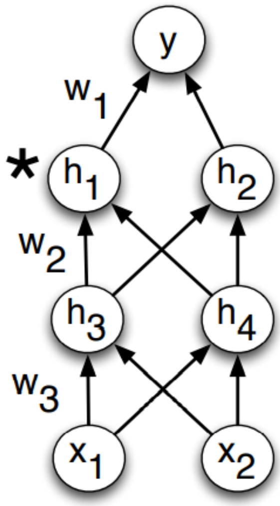


Figure 3: Neural Network with four layers. (Note: Each of $w1$, $w2$, and $w3$ refers to the weight on a single connection, not the whole layer.)

1. **[3 points]** $\partial C/\partial w_1 = 0$: ____, your explanation:

> **Solution**
> TRUE.
> Because h_1 is zero, and therefore changing w_1 doesn't affect the input to unit y. Therefore it doesn't affect the output of the network, or the cost.

2. **[3 points]** $\partial C/\partial w_2 = 0$: ____, your explanation:

> **Solution**
> TRUE.
> Because the input z_1 is negative, $\partial h_1/\partial z_1 = 0$, so changing w_2 by a small amount doesn't change h_1. Therefore it has no effect on the output of the network.

3. **[3 points]** $\partial C/\partial w_3 = 0$: ____, your explanation:

> **Solution**
> FALSE.
> Changing w_3 by a small amount can change h_3, which can change h_2, which can change y, which can change C.

IX Convex Sets and Convex Functions [12 points]

In this problem, you should first write down whether the set or the function is convex or non-convex, then either prove the set or the function is convex or provide an example to show that it's non-convex.
Note: correct answers without proof will get half the points.

1. **[6 points]** Determine the convexity of the following sets:

   (a) Polyhedra:
   $$\{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{x} \preceq \mathbf{b}, \mathbf{C}\mathbf{x} = \mathbf{d}\},$$
   where $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{b} \in \mathbb{R}^m$, $\mathbf{C} \in \mathbb{R}^{p\times n}$, and $\mathbf{d} \in \mathbb{R}^p$.

   Solution
   Polyhedra is the intersection of m halfspaces $\mathbf{A}\mathbf{x} \preceq \mathbf{b}$ and p hyperplanes $\mathbf{C}\mathbf{x} = \mathbf{d}$.
   It is convex because:
   - Both halfspaces and hyperplanes are convex sets;
   - Intersection preserves convexity.

   (b) Positive semidefinite cone:
   $$\mathbb{S}^n_+ = \{\mathbf{X} \in \mathbb{S}^n \mid \mathbf{X} \succeq 0\},$$
   where $\mathbb{S}^n$ denotes the set of symmetric matrices in $\mathbb{R}^{n\times n}$. Here $\mathbf{X} \succeq 0$ represents the generalized inequality on matrices, indicating $\mathbf{z}^\top\mathbf{X}\mathbf{z} \geq 0$, $\forall \mathbf{z} \in \mathbb{R}^n$.

   Solution
   Given $\mathbf{X} \in \mathbb{S}^n_+$ and $\mathbf{Y} \in \mathbb{S}^n_+$, we need to prove that $\theta\mathbf{X} + (1-\theta)\mathbf{Y} \in \mathbb{S}^n_+$.
   $$\begin{aligned} &\mathbf{z}^\top\left(\theta\mathbf{X} + (1-\theta)\mathbf{Y}\right)\mathbf{z} && (14)\\ &= \theta\mathbf{z}^\top\mathbf{X}\mathbf{z} + (1-\theta)\mathbf{z}^\top\mathbf{Y}\mathbf{z} && (15)\\ &\leq \theta 0 + (1-\theta)0 && (16)\\ &= 0 && (17)\end{aligned}$$
   where $\mathbf{z} \in \mathbb{R}^n$ is arbitrary.

2. **[6 points]** Determine the convexity of the following functions:

   (a) Lasso objective:
   $$f(\mathbf{x}) = ||\mathbf{A}\mathbf{x} - \mathbf{b}||^2 + \lambda||\mathbf{x}||_1,$$
   where $\mathbf{A} \in \mathbb{R}^{m\times n}$ ($\mathbf{A}^\top\mathbf{A} \in \mathbb{S}^n_+$), $\mathbf{b} \in \mathbb{R}^m$, $\lambda > 0$, and $||\cdot||$ and $||\cdot||_1$ denote ℓ_2-norm and ℓ_1-norm, respectively.

   Solution
   Let $f(x) = g(x) + h(x)$, with $g(x) = ||\mathbf{A}\mathbf{x} - \mathbf{b}||^2$ and $h(x) = \lambda||\mathbf{x}||_1$.
   - Since $\nabla^2 g = \mathbf{A}^\top\mathbf{A} \succeq 0$, $g(x)$ is convex;
   - Because $||\mathbf{x}||_1$ is convex and nonnegative multiple preserves convexity, $h(x)$ is convex;
   - Because summation preserves convexity, $g(x) + h(x)$ is convex.

   Hence, $f(x)$ is convex.

   (b) Weighted log barrier for linear inequalities:
   $$f(\mathbf{x}) = -\sum_{i=1}^{m} c_i \log(b_i - \mathbf{a}_i^\top\mathbf{x}),$$
   with $\mathbf{dom}f = \{\mathbf{x} \mid \mathbf{a}_i^\top\mathbf{x} < b_i,\ i = 1, 2, ..., m\}$. Here $\mathbf{a}_i, \mathbf{x} \in \mathbb{R}^n$, and $c_i > 0$ denotes the weighting coefficient, $i = 1, 2, ..., m$.

Solution

We can rewrite $f(x) = \sum_{i=1}^m c_i \left(-\log(b_i - \mathbf{a}_i^\top \mathbf{x})\right)$.

- Because $-\log(x)$ is convex and composition with affine function preserves convexity, $-\log(b_i - \mathbf{a}_i^\top \mathbf{x})$ is convex;
- Because nonnegative multiple preserves convexity, $c_i \left(-\log(b_i - \mathbf{a}_i^\top \mathbf{x})\right)$;
- Because summation preserves convexity, $\sum_{i=1}^m c_i \left(-\log(b_i - \mathbf{a}_i^\top \mathbf{x})\right)$ is convex.

Hence, $f(x)$ is convex.