Quiz 1

March 18th 2020

1 Lecture 5

$$
\begin{aligned}
\log \frac{Pr(G=1|X=x)}{1-Pr(G=1|X=x)} &= \beta_0 + x^T\beta \\
\frac{Pr(G=1|X=x)}{1-Pr(G=1|X=x)} &= \exp(\beta_0 + x^T\beta) \\
Pr(G=1|X=x) &= \frac{\exp(\beta_0 + x^T\beta)}{1+\exp(\beta_0 + x^T\beta)} \\
Pr(G=2|X=x) &= 1 - Pr(G=1|X=x) \quad = \frac{1}{1+\exp(\beta_0 + x^T\beta)}
\end{aligned}
$$

2 Lecture 6

$$
\begin{aligned}
\hat{\Sigma}* &= \frac{\sum_{k=1}^{K}\sum_{g_i=k}(x_i^* - \hat{\mu}_k^*)(x_i^* - \hat{\mu}_k^*)^T}{N-K} \\
&= \frac{\sum_{k=1}^{K}\sum_{g_i=k}(\hat{\Sigma}^{-\frac{1}{2}}x_i - \hat{\Sigma}^{-\frac{1}{2}}\hat{\mu}_k)(\hat{\Sigma}^{-\frac{1}{2}}x_i - \hat{\Sigma}^{-\frac{1}{2}}\hat{\mu}_k)^T}{N-K} \\
&= \frac{\sum_{k=1}^{K}\sum_{g_i=k}\hat{\Sigma}^{-\frac{1}{2}}(x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T\hat{\Sigma}^{-\frac{1}{2}}}{N-K} \\
&= \hat{\Sigma}^{-\frac{1}{2}}\frac{\sum_{k=1}^{K}\sum_{g_i=k}\hat{(}x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T}{N-K}\hat{\Sigma}^{-\frac{1}{2}} \\
&= \hat{\Sigma}^{-\frac{1}{2}}\hat{\Sigma}\hat{\Sigma}^{-\frac{1}{2}} \\
&= I
\end{aligned}
$$

Solutions to Quizzes in Lectures 7 and 8

Lu Sun

March 30, 2020

1 Solution to Quiz in Lecture 7

1.1 Probability Density Function

Suppose that we have a categorical random variable X with K states, i.e., $X \in \{1, 2, ..., K\}$. Let θ_k denote the probability of $X = k$ ($k = 1, 2, ..., K$), the probability density function is defined by

$$P(X|\theta) = \theta_1^{\mathbf{1}_{X=1}} \theta_2^{\mathbf{1}_{X=2}} \cdots \theta_K^{\mathbf{1}_{X=K}}, \tag{1}$$

where $\theta = \{\theta_1, \theta_2, ..., \theta_K\}$, and $\mathbf{1}_{(\cdot)}$ is the indicator function.

1.2 Likelihood Function

Given a training dataset $\mathcal{D} = \{x_1, x_2, ..., x_N\}$, in which each sample x_i is an observation of X, the likelihood function becomes

$$\begin{aligned} L(\theta) &= P(\mathcal{D}|\theta) \\ &= P(x_1, x_2, ..., x_N|\theta) \\ &= \prod_{i=1}^{N} P(x_i|\theta) \\ &= \prod_{i=1}^{N} \theta_1^{\mathbf{1}_{x_i=1}} \theta_2^{\mathbf{1}_{x_i=2}} \cdots \theta_K^{\mathbf{1}_{x_i=K}} \\ &= \theta_1^{\sum_{i=1}^{N} \mathbf{1}_{x_i=1}} \theta_2^{\sum_{i=1}^{N} \mathbf{1}_{x_i=2}} \cdots \theta_K^{\sum_{i=1}^{N} \mathbf{1}_{x_i=K}} \\ &= \theta_1^{\alpha_1} \theta_2^{\alpha_2} \cdots \theta_K^{\alpha_K}, \end{aligned} \tag{2}$$

where α_k denotes the number of $X = k$ in the training dataset $\mathcal{D}$, thus $\alpha_k = \sum_{i=1}^{N} \mathbf{1}_{x_i=k}$, $\forall k$.

1.3 Prior Probability

If the prior of θ are from the Dirichlet$(\beta_1, \beta_2, ..., \beta_K)$, we have

$$P(\theta) = \frac{\theta_1^{\beta_1-1} \theta_2^{\beta_2-1} \cdots \theta_K^{\beta_K-1}}{B(\beta_1, \beta_2, ..., \beta_K)}. \tag{3}$$

In (3), β_k ($\forall k$) is the hyperparameter of Dirichlet distribution, and $B(\cdot)$ denotes the beta distribution, that is irrelevant with θ.

1.4 Posterior Probability

By combining (2) and (3), log-posterior is formulated as follows:

$$\begin{aligned}\ln P(\theta|\mathcal{D}) &\propto \ln\left(P(\mathcal{D}|\theta)P(\theta)\right)\\ &\propto \ln\left(\theta_1^{\alpha_1+\beta_1-1}\theta_2^{\alpha_2+\beta_2-1}\cdots\theta_K^{\alpha_K+\beta_K-1}\right)\\ &\propto \sum_{k=1}^{K}(\alpha_k+\beta_k-1)\ln\theta_k. \end{aligned}\tag{4}$$

Based on the fact that $\sum_{k=1}^{K}\theta_k=1$, there are $K-1$ independent parameters in $\{\theta_1,\theta_2,...,\theta_K\}$. Thus we can treat $\theta_K=1-\sum_{k=1}^{K-1}\theta_k$ as the dependent parameter. As the log-posterior is a concave function w.r.t. θ, its global maximum is obtained by setting its derivative equal to 0, leading to

$$\begin{aligned}\frac{\partial\ln P(\theta|\mathcal{D})}{\partial\theta_k} &= \frac{\alpha_k+\beta_k-1}{\theta_k}-\frac{\alpha_K+\beta_K-1}{1-\sum_{k=1}^{K-1}\theta_k}\\ &= \frac{\alpha_k+\beta_k-1}{\theta_k}-\frac{\alpha_K+\beta_K-1}{\theta_K}\\ &= 0.\end{aligned}\tag{5}$$

Obviously,

$$\hat{\theta}_k=\frac{\alpha_k+\beta_k-1}{\alpha_K+\beta_K-1}\hat{\theta}_K.\tag{6}$$

Substituting (6) into $\sum_{k=1}^{K}\theta_k=1$, gives rise to

$$\hat{\theta}_K=\frac{\alpha_K+\beta_K-1}{\sum_{k=1}^{K}\alpha_k+\beta_k-1}.\tag{7}$$

By combing (6) and (7), we reach our conclusion:

$$\hat{\theta}_k=\frac{\alpha_k+\beta_k-1}{\sum_{k=1}^{K}\alpha_k+\beta_k-1},\quad k=1,2,...,K.\tag{8}$$

2 Solution to Quiz in Lecture 8

The solution is the MLE version of the above one, by replacing X and θ by Y and π, respectively.

SI 151
The solution of quiz 5

Xin Deng

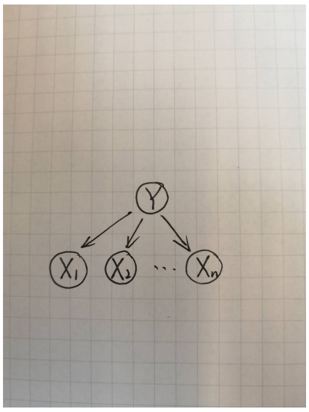
April 2, 2020

1. What is the Bayes Network of the diagonal LDA?

   Solution:
   According to the slide of lecture 6, we have

$$\begin{aligned} P(Y|X) &\propto P(X,Y) \\ &= P(X|Y) \cdot P(Y) \\ &= P(Y) \cdot \prod_i P(X_i|Y) \end{aligned}$$

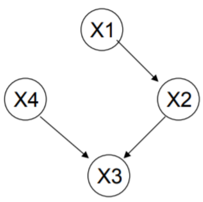
   Then the Bayes Network of diagonal LDA is given as



2. Use the D-separation to analyze the following cases:

   (a) X_1 and X_4 are conditionally independent given $\{X_2, X_3\}$.

   (b) X_1 and X_4 are not conditionally independent given X_3.



   Solution:

   (a) From X_2 to X_4, it's the head to head situation. Then X_2 and X_4 are not conditionally independent given X_3. But given X_2, the path from X_3 to X_1 is blocked according to the head to tail situation. Therefore, the statement (a) is true.

   (b) It is similar to the analysis of statement (a). The path from X_4 to X_2 is open given X_3 according to the head to head situation. Further, it is also unblocked from X_3 to X_1. Therefore, X_1 and X_4 are not conditionally independent given X_3.

Quiz 6

Yuyan Zhou

April 10, 2020

1. Initialize $\boldsymbol{\theta}$
2. Repeat
3. E-step: Use $\boldsymbol{X}$ and current $\boldsymbol{\theta}$ to calculate $P(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})$
4. M-step: Replace current $\boldsymbol{\theta}$ by

$$\boldsymbol{\theta} \leftarrow arg\max_{\boldsymbol{\theta'}} Q(\boldsymbol{\theta'}|\boldsymbol{\theta}) + logP(\boldsymbol{\theta'})$$

where $Q(\boldsymbol{\theta'}|\boldsymbol{\theta}) = E_{P(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}[logP(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta'})]$
5. Until convergence

Only M-step is changed, because in MAP, we have

$$\begin{aligned}
&E_{P(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}[log(P(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta'})P(\boldsymbol{\theta'}))] \\
&= E_{P(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}[logP(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta'})] + E_{P(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}[logP(\boldsymbol{\theta'})] \\
&= E_{P(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}[logP(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta'})] + logP(\boldsymbol{\theta'}) \\
&= Q(\boldsymbol{\theta'}|\boldsymbol{\theta}) + logP(\boldsymbol{\theta'})
\end{aligned}$$

Reference Solution to the Quiz 7

Xiangyu Yang

April 15, 2020

1 Lecture 13

According to Theorem 7.1 shown in the course slide, please derive the following sample complexity for the consistent learner, which reads

$$m \geq \frac{1}{\epsilon}\left[\ln(|H|) + \ln(\frac{1}{\delta})\right]. \tag{1}$$

Proof. By Theorem 7.1, and let $\delta > 0$ be an upper bound on the probability of not exhausting the version space, so

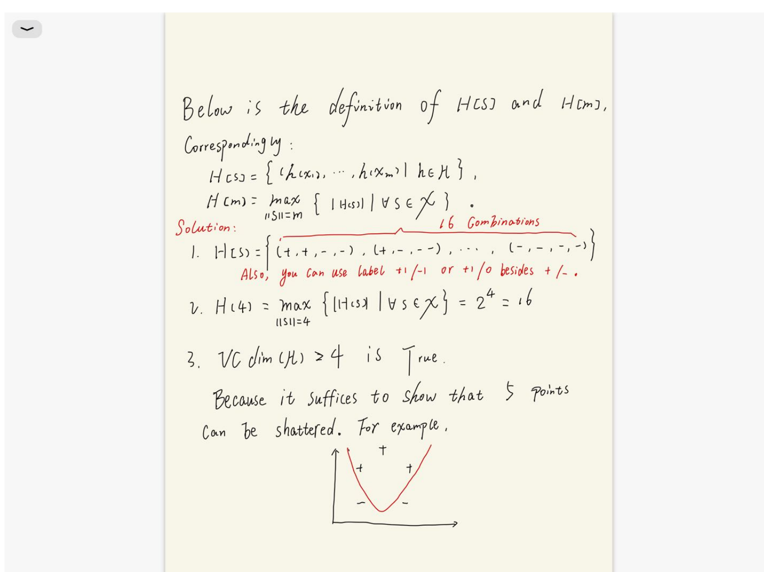
$$\Pr(\exists h \in VS_{H,D}, err_D(h) \geq \epsilon) \leq |H|e^{-\epsilon m} \leq \delta. \tag{2}$$

Focus on the second inequality of (2), we have

$$|H|e^{-\epsilon m} \leq \delta \Longleftrightarrow \ln |H|e^{-\epsilon m} \leq \ln \delta. \tag{3}$$

Hence, after some simple algebraic manipulations, we can easily obtain the desired inequality (1). This completes the proof. □

2 Lecture 14



Below is the definition of $H[S]$ and $H[m]$, Correspondingly:

$H[S] = \{(h(x_1), \cdots, h(x_m)) \mid h \in \mathcal{H}\}$,

$H(m) = \max_{\|S\|=m} \{ |H(S)| \mid \forall S \in \mathcal{X} \}$.

Solution:

1. $H[S] = \{ \overbrace{(+,+,-,-), (+,-,-,-), \cdots, (-,-,-,-)}^{16 \text{ Combinations}} \}$

Also, you can use label +1/-1 or +1/0 besides +/-.

2. $H(4) = \max_{\|S\|=4} \{ |H(S)| \mid \forall S \in \mathcal{X} \} = 2^4 = 16$

3. $VC\dim(\mathcal{H}) \geq 4$ is True.

Because it suffices to show that 5 points Can be shattered. For example,

Quiz 1

March 18th 2020

1 Lecture 15

$$
\begin{aligned}
f &= (1-\epsilon_t)e^{-\alpha} + \epsilon_t e^{\alpha} \\
\nabla f &= 0 \\
-(1-\epsilon_t)e^{-\alpha} + \epsilon_t e^{\alpha_t} &= 0 \\
\alpha_t &= \frac{1}{2}\log(\frac{1-\epsilon_t}{\epsilon_t})
\end{aligned}
$$

2 Lecture 16

1. AdaBoost increases the margins
2. Large margin in training indicates lower generalization error, independent of the number of rounds of boosting.

Solution

April 30, 2020

Lecture 17

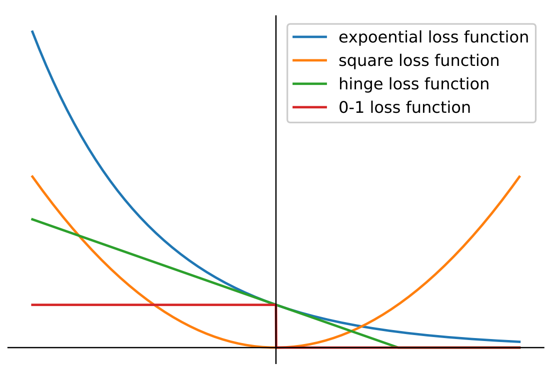
$$\begin{cases} \gamma_1 \frac{w}{\|w\|} = x_1 - x_0 \\ w^\top x_0 = 0 \\ w^\top x_1 = 1 \end{cases}$$

$$\Longrightarrow \gamma_1 \frac{w^\top w}{\|w\|} = w^\top x_1 - w^\top x_0 = 1$$

$$\Longrightarrow \gamma_1 \frac{\|w\|^2}{\|w\|} = 1$$

$$\Longrightarrow \gamma_1 = \frac{1}{\|w\|}$$

Lecture 18



SI 151
The solution of quiz 10

Xin Deng

May 7, 2020

1. What is the difference between semi-supervised learning and active learning ?

   Solution:
   In semi-supervised learning, the date which experts need to label are sampled randomly. While in active learning, we sample the data based on Active Query, i.e., some sampling rules.

Quiz for lecture 21 and 22

Yuyan Zhou

May 13, 2020

1 lecture 21

$$
\begin{aligned}
\mu &= \frac{1}{n}\sum_{1}^{n} x^i \\
\frac{1}{n}\sum_{1}^{n} \|x^i - c\|^2 &= \frac{1}{n}\sum_{1}^{n} \|x^i - \mu + \mu - c\|^2 \\
&= \frac{1}{n}\sum_{1}^{n} \|x^i - \mu\|^2 + \frac{1}{n}\sum_{1}^{n} \|\mu - c\|^2 + \frac{2}{n}\sum_{1}^{n}(x^i - \mu)^T(\mu - c) \\
&= \frac{1}{n}\sum_{1}^{n} \|x^i - \mu\|^2 + \|\mu - c\|^2 + 0^T(\mu - c) \\
&= \frac{1}{n}\sum_{1}^{n} \|x^i - \mu\|^2 + \|\mu - c\|^2
\end{aligned}
$$

take derivative w.r.t c and set it to 0, then we have the optimal $c = \mu$

2 lecture 22

Each image except the top left one forms an eigenvector, so there are 15 eigenvectors in total.
We can reconstruct the image by the following steps:

1. Reshape each image as a "long" vector $v_i,\ i \in \{1, \dots, 15\}$ and x
2. calculate the coefficient by projecting x onto each v_i, and we get $\langle x, v_i \rangle$
3. construct a linear combination $\hat{x} = \sum_1^{15} \langle x, v_i \rangle v_i$
4. reshape $\hat{x}$ back to the matrix shape

Reference Solutions to the Quiz 7

Xiangyu Yang

May 21, 2020

1 Lecture 23

1). Please derive the updating rule, if we use ReLU as the activation function.

Sol: Before proceeding, we introduce the indicator function $\mathbb{I}(\cdot)$, meaning if condition $\cdot$ is met, then return 1; otherwise, return 0.

In the backward pass, we consider changes in any w_i, $i = 1, \ldots, n$ affecting the total error E. This is achieved by simply applying the chain rule, i.e.,

$$\begin{aligned} \frac{\partial E}{\partial w_i} &= \sum_d \frac{\partial E}{\partial o_d} \frac{\partial o_d}{\partial \text{net}_d} \frac{\partial \text{net}_d}{\partial w_i} \\ &= \sum_d (o_d - t_d) \mathbb{I}(\text{net}_d \geq 0) x_{d,i}, \end{aligned} \tag{1}$$

where we use the fact that the derivative of ReLU function is the defined indicator function above. We hence update the weights as follows

$$\begin{aligned} w_i &= w_i - \eta \frac{\partial E}{\partial w_i} \\ &= w_i - \eta \sum_d (o_d - t_d) \mathbb{I}(\text{net}_d \geq 0) x_{d,i}. \end{aligned} \tag{2}$$

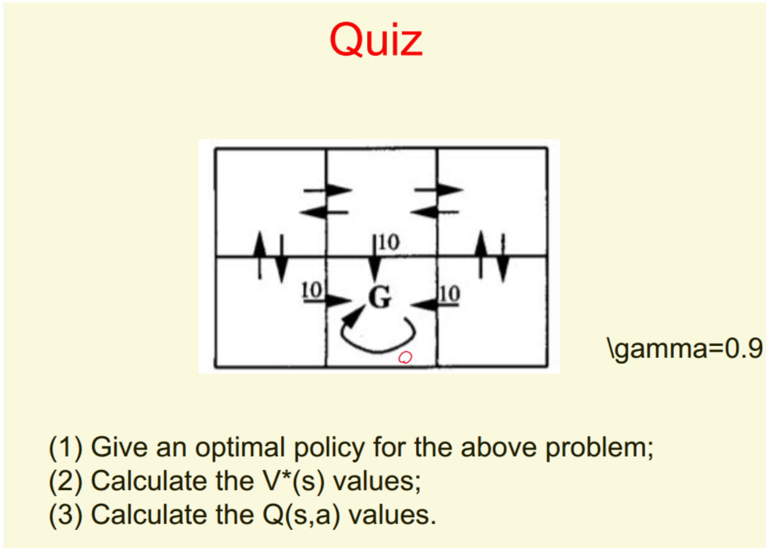
2). Compare the difference between the error gradients of the sigmoid function and the ReLU function.

Sol: The error gradients of the sigmoid function reads

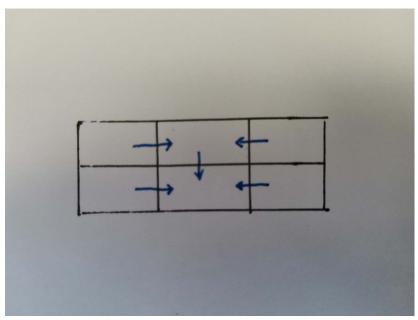
$$\begin{aligned} \frac{\partial E}{\partial w_i} &= \sum_d \frac{\partial E}{\partial o_d} \frac{\partial o_d}{\partial \text{net}_d} \frac{\partial \text{net}_d}{\partial w_i} \\ &= \sum_d (o_d - t_d) o_d (1 - o_d) x_{d,i}. \end{aligned} \tag{3}$$

We first note that the backward updating is a gradient-based learning method. From (3), we observe that the derivative of the sigmoid function is always smaller than 1 (i.e., consider $o_d(1 - o_d)$). Indeed, it is at most 0.25. This would cause significant side effects if you have many layers as the product of many smaller than 1 values goes to zero very quickly. However, RELU activation fixes the vanishing gradients problem because it only saturates in one direction.

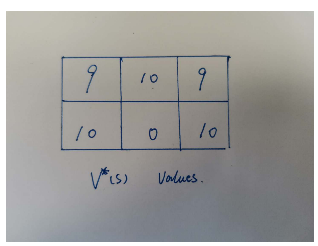
2 Lecture 24



Sol:

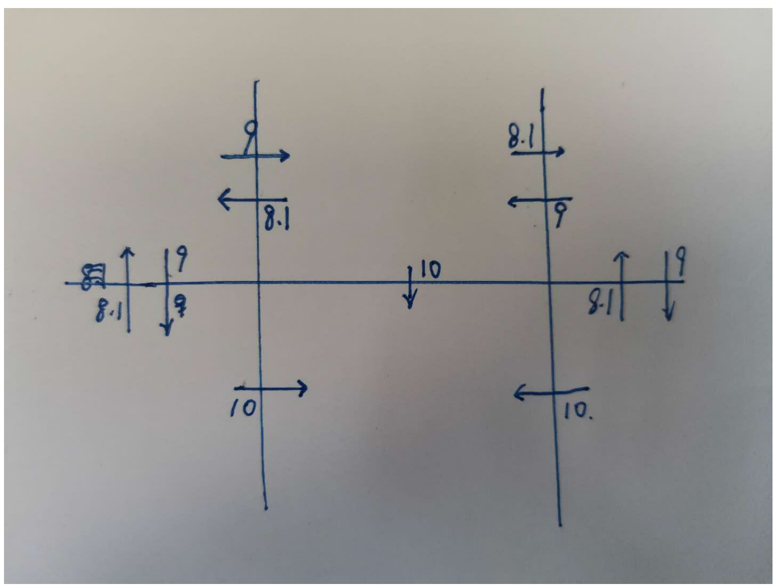


(1)



(2)

(3)


9
8.1
8.1
9
9
8.1
9
10
8.1
9
10
10.

Week 13 Quiz

March 18th 2020

1 Lecture 25

Given $X, Y \in \mathbb{S}^n_{++}$, $\forall z \in \mathbb{R}^n, z^T X z > 0, z^T Y z > 0$.
For $\theta_1, \theta_2 \geq 0$, then $\forall z$,

$$\begin{aligned} z^T(\theta_1 X + \theta_2 Y)z &= \theta_1 z^T X z + \theta_2 z^T Y z \\ &\geq 0 + 0 \\ &= 0 \end{aligned}$$

If $\theta_1 = \theta_2 = 0$, $z^T(\theta_1 X + \theta_2 Y)z = 0 \notin \mathbb{S}^n_{++}$. So $\mathbb{S}^n_{++}$ is not a convex cone.

2 Lecture 26

$\forall Y_1, Y_2 \in C$, we can get $\forall \theta \in (0,1), \theta Y_1 + (1-\theta)Y_2 \in C$.
$\forall x_1, x_2 \in f^{-1}(C), f(x_1) = X_1, f(x_2) = X_2$, and $\forall \theta \in (0,1)$

$$\begin{aligned} f(\theta x_1 + (1-\theta)x_2) &= A(\theta x_1 + (1-\theta)x_2) + b \\ &= \theta A x_1 + \theta b + (1-\theta)A x_2 + (1-\theta)b \\ &= \theta f(x_1) + (1-\theta) f(x_2) \\ &= \theta X_1 + (1-\theta)X_2 \\ &\in C \end{aligned}$$

So $f^{-1}(C)$ is convex.

Quiz Solutions

June 7, 2020

Lecture 27

$$g\left(\theta x_1 + (1-\theta)x_2\right) = \sup_{y\in A} f\left(\theta x_1 + (1-\theta)x_2, y\right).$$

Since $f(x,y)$ is convex x, we have

$$\begin{aligned}
\sup_{y\in A} f\left(\theta x_1 + (1-\theta)x_2, y\right) &\le \sup_{y\in A} \theta f\left(x_1, y\right) + \sup_{y\in A}(1-\theta) f\left(x_2, y\right) \\
&\le \theta \sup_{y\in A} f\left(x_1, y\right) + (1-\theta)\sup_{y\in A} f\left(x_2, y\right) \\
&= \theta g\left(x_1\right) + (1-\theta) g\left(x_2\right)
\end{aligned}$$

Therefore,

$$g\left(\theta x_1 + (1-\theta)x_2\right) \le \theta g\left(x_1\right) + (1-\theta) g\left(x_2\right),$$

namely, $g(x)$ is convex.

Lecture 28

Here, we consider the following standard Gaussian distribution, i.e., $\mu = 0, \sigma = 1$,

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Recall that f is log-concave if and only if $f''(x)f(x) \le f'(x)^2$ for all x. We first calculate $f''(x)$ and $f'(x)$,

$$\begin{aligned}
f'(x) &= -\frac{1}{\sqrt{2\pi}} e^{-x^2/2} x = -f(x)x \\
f''(x) &= \frac{1}{\sqrt{2\pi}} e^{-x^2/2} x^2 - \frac{1}{\sqrt{2\pi}} e^{-x^2/2} = f(x)x^2 - f(x).
\end{aligned}$$

Clearly,

$$f''(x)f(x) = f(x)^2(x^2 - 1) \leq f(x)^2 x^2 = f'(x)^2,$$

which implies $f(x)$ is log-concave. The result can be readily generalized for any μ and σ.

SI 151, Spring 2020
The solution of quiz 15

1. **Solution:**
A quadratic program can be expressed in the form

$$\begin{array}{ll} \text{minimize}_x & \frac{1}{2}x^TQx + r^Tx + s \\ \text{subject to} & Gx \preceq h, \\ & Ax = b, \end{array}$$

where $Q \in \mathbb{S}_+^n, G \in \mathbb{R}^{m\times n}$ and $A \in \mathbb{R}^{p\times n}$. The original QP can be rewritten in epigraph form as the following QP in $x \in \mathbb{R}^n$ and $t \in \mathbb{R}$

$$\begin{array}{ll} \text{minimize}_t & t \\ \text{subject to} & \frac{1}{2}x^TQx + r^Tx + s \leq t, \\ & Gx \preceq h, \\ & Ax = b. \end{array}$$

Since Q is symmetric and positive semidefinite, there is some matrix P such that

$$Q = P^TP.$$

Using the Schur complement, the convex quadratic inequality constraint can be rewritten as the following LMI

$$\begin{bmatrix} -I & -Px \\ -x^TP^T & -t+s+r^Tx \end{bmatrix} \preceq 0$$

and the linear inequality constraint can be written as the following LMI

$$\mathbf{diag}(Gx-h) \preceq 0.$$

Thus, the convex QP can be written as the SDP in $x \in \mathbb{R}^n$ and $t \in \mathbb{R}$

$$\begin{array}{ll} \text{minimize}_{x,t} & t \\ \text{subject to} & \begin{bmatrix} -I & -Px & 0 \\ -x^TP^T & -t+s+r^Tx & 0 \\ 0 & 0 & \mathbf{diag}(Gx-h) \end{bmatrix} \preceq 0 \end{array}$$

2. **Solution:**
The Lagrangian is

$$L(x, z, \mu) = \sum_{i=1}^n x_i \log x_i + \lambda^T(Ax-b) + \mu^T(Cx-d).$$

Minimizing over x_i gives the conditions

$$1 + \log x_i + a_i^T\lambda + c_i^T\mu = 0, \quad i = 1, ..., n,$$

with solution

$$x_i = e^{-a_i^T\lambda - c_i^T\mu - 1},$$

where a_i and c_i are the ith column of A and C, respectively. Plugging this in in L gives the Lagrange dual function

$$g(\lambda, \mu) = -b^T\lambda - d^T\mu - \sum_{i=1}^n e^{-a_i^T\lambda - c_i^T\mu - 1}.$$