EE488C Special Topics in EE <Deep Learning and AlphaGo> 1:00-3:00pm, October 24, 2016
EE, KAIST, Fall 2016

Mid-term exam solutions

---

1. (10 points) *Find an example of two binary random variables X and Y such that X and Y are Bernoulli(1/2) and they are not independent.*

Solution) If $X = Y$ is Bernoulli(1/2), then X and Y are not independent since $\Pr(X = 0)\Pr(Y = 0) = \frac{1}{4} \neq \Pr(X = 0, Y = 0) = \frac{1}{2}$.

2. (20 points) *Consider a neural network given as follows.*

$$y = \mathbf{w}^T\mathbf{h},\ \mathbf{h} = g(\mathbf{W}^T\mathbf{x})$$

where $\mathbf{x} = (x_1, x_2)^T$ *is the input vector, y is the output,* $\mathbf{W} = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix}$, $\mathbf{w} = (w_1, w_2)^T$, *and $g(\cdot)$ is an elementwise ReLU. Assume* $\mathbf{W} = \begin{pmatrix} \alpha & \beta \\ \beta & \alpha \end{pmatrix}$ *and* $\mathbf{w} = (\gamma, \gamma)^T$ *and find conditions on α, β and γ with which this network can perform XOR, i.e., the output is 0 if the input is $(0,0)^T$ or $(1,1)^T$ and the output is 1 if the input is $(0,1)^T$ or $(1,0)^T$.*

Solution) As done in lecture notes, let's define $\mathbf{X} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix}$. Using this, we get $\mathbf{XW} = \begin{pmatrix} 0 & 0 \\ \beta & \alpha \\ \alpha & \beta \\ \alpha+\beta & \alpha+\beta \end{pmatrix}$ and thus $g(\mathbf{XW})\mathbf{w} = \begin{pmatrix} 0 \\ \gamma(g(\alpha)+g(\beta)) \\ \gamma(g(\alpha)+g(\beta)) \\ 2\gamma g(\alpha+\beta) \end{pmatrix}$, which needs to be equal to $\begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \end{pmatrix}$.
Since $\gamma(g(\alpha)+g(\beta)) = 1$ and $g(\cdot) \geq 0$, we need to have $\gamma > 0$ and $g(\alpha)+g(\beta) > 0$, which implies $\alpha > 0$ or $\beta > 0$. Furthermore, $\alpha + \beta \leq 0$ because $2\gamma g(\alpha+\beta) = 0$ and $\gamma > 0$. Therefore, all possible values for α, β and γ are

$$\alpha > 0,\ \gamma = \frac{1}{\alpha},\ \beta \leq -\alpha$$

or

$$\beta > 0,\ \gamma = \frac{1}{\beta},\ \alpha \leq -\beta.$$

3. (20 points) *Consider a problem of estimating θ from a random sample x generated according to Bernoulli(θ), where $0 \leq \theta \leq 1$. Let's consider a trivial estimator that always says α regardless of the sample it observes, where $0 \leq \alpha \leq 1$.*

(a) *Calculate the MSE of this estimator.*

Solution) $\mathbb{E}[(\alpha-\theta)^2] = (\alpha-\theta)^2$

(b) *Find the value of α that minimizes the worst-case MSE, i.e., find α that minimizes the following. What is the minimum worst-case MSE?*

$$\max_{0\leq\theta\leq 1} \mathbb{E}[(\alpha-\theta)^2]$$

Solution) Since $\mathbb{E}[(\alpha-\theta)^2] = (\alpha-\theta)^2$, it is maximizes either at $\theta = 0$ or $\theta = 1$, i.e.,

$$\max_{0\le\theta\le 1} \mathbb{E}[(\alpha-\theta)^2] = \max\{\alpha^2, (1-\alpha)^2\}$$
$$= \begin{cases} \alpha^2 & \text{if } \alpha \ge \frac{1}{2} \\ (1-\alpha)^2 & \text{otherwise} \end{cases}$$

Therefore, this is minimized if $\alpha = \frac{1}{2}$ and the minimum worst-case MSE is equal to $\frac{1}{4}$. You can see this is higher than $\frac{1}{16}$ achievable by the optimal estimator in Problem 1(b) in Problem Set # 2.

4. (20 points) *Let $J(x,y) = (x-2)^2 + 5(y-2)^2$ denote a cost function in terms of $(x,y) \in \mathbb{R}$. Let's add an L^1 regularization term $5\|(x,y)\|_1 = 5(|x|+|y|)$ and minimize $J(x,y) + 5(|x|+|y|)$. We can minimize this by following the steps below. The case $y \le 0$ can be analyzed similarly, but the minimum of $J(x,y) + 5(|x|+|y|)$ is achieved when $y \ge 0$ and therefore we only need to analyze the following two cases to find the minimum value of $J(x,y) + 5(|x|+|y|)$.*

(a) *Assume $x \ge 0$ and $y \ge 0$ and minimize $J(x,y) + 5(|x|+|y|)$.*

Solution) Since $x \ge 0$ and $y \ge 0$, we have $|x| = x$ and $|y| = y$. Therefore, we get

$$J(x,y) + 5(|x|+|y|) = (x-2)^2 + 5(y-2)^2 + 5(x+y) = (x+1/2)^2 + 5(y-3/2)^2 + \text{const}$$

This is minimized if $y = \frac{3}{2}$ and $x = 0$. Note that since $x \ge 0$, $(x+1/2)^2$ is minimized when $x = 0$. The minimum value is $J(0,3/2) + 5(0+3/2) = \frac{51}{4}$.

(b) *Assume $x \le 0$ and $y \ge 0$ and minimize $J(x,y) + 5(|x|+|y|)$.*

Solution) Since $x \le 0$ and $y \ge 0$, we have $|x| = -x$ and $|y| = y$. Therefore, we get

$$J(x,y) + 5(|x|+|y|) = (x-2)^2 + 5(y-2)^2 + 5(-x+y) = (x-9/2)^2 + 5(y-3/2)^2 + \text{const}$$

This is minimized if $y = \frac{3}{2}$ and $x = 0$. Note that since $x \le 0$, $(x-9/2)^2$ is minimized when $x = 0$. The minimum value is $J(0,3/2) + 5(0+3/2) = \frac{51}{4}$. Therefore, by combining (a) and (b) we conclude that the minimum of $J(x,y) + 5(|x|+|y|)$ achieved when $x = 0$ and $y = \frac{3}{2}$ and its minimum value is $\frac{51}{4}$. You can see this gives a sparse solution.

5. (10 points) *Consider the following one-tap IIR filter for averaging $u[k]$, where $0 < \alpha < 1$.*

$$x[k] = \alpha x[k-1] + (1-\alpha)u[k]$$

Assume $u[k] = \begin{cases} 1 & \text{if } k = 0 \\ 0 & \text{otherwise} \end{cases}$ and $x[k] = 0$ for $k < 0$. Calculate $\sum_{k=0}^{\infty} kx[k]$, which can be considered as an effective width of the impulse response.

Solution) Since $x[k] = (1-\alpha)\alpha^k$ for $k \geq 0$ and $x[k] = 0$ for $k < 0$, we get

$$\begin{aligned}
\sum_{k=0}^{\infty} kx[k] &= \sum_{k=0}^{\infty} k(1-\alpha)\alpha^k \\
&= \alpha(1-\alpha)\sum_{k=0}^{\infty} k\alpha^{k-1} \\
&= \alpha(1-\alpha)\frac{\partial}{\partial \alpha}\sum_{k=0}^{\infty} \alpha^k \\
&= \alpha(1-\alpha)\frac{\partial}{\partial \alpha}\frac{1}{1-\alpha} \\
&= \alpha(1-\alpha)\frac{1}{(1-\alpha)^2} \\
&= \frac{\alpha}{1-\alpha}.
\end{aligned}$$

6. (20 points) *Design a continuous function $f : \mathbb{R}^2 \to \mathbb{R}$ such that*

$$\begin{aligned}
f(x,0) &= 0, \ \forall x \in \mathbb{R} \\
f(0,y) &= 0, \ \forall y \in \mathbb{R} \\
f(1,1) &= 1 \\
f(-1,-1) &= 1
\end{aligned}$$

Try to make $f(x,y)$ as simple as you can such that it can generalize well for unseen samples.

Solution) $f(x,y) = xy$ would be a simple solution. You can see this satisfies the conditions given above. Of course, there are infinitely many solutions, e.g., $f(x,y) = xy + (xy-1)^2$. But, as the function becomes more complex, it tends to overfit and may perform poorly for unseen samples.