

در تمامی مسائلی که لازم است یک گزینه را انتخاب کنید، محتمل‌ترین گزینه که یک مهندس داده آن را برمی‌گزیند، انتخاب نمایید.

$ReLU(x)$	$\tanh(x)$	$Sigmoid(x)$	$ReLU(-x)$	$Sigmoid(-x)$	$\exp(x^2)$	
x	✓	x	x	x	x	پاسخ بند الف
x	x	x	✓	✓	x	پاسخ بند ب
x	x	x	x	✓	x	پاسخ بند ج

الف) با توجه به مقایسه خروجی با مقادیر هدف +۱ و -۱، (ب) لازم است در صورت عدم تطابق علامت هزینه افزایش یابد و برعکس، (ج) در مقادیر بزرگ اشباع شود.

سوال دو-۱۲ (نمره)

روش K-NN	نمی‌تواند، ماهیت instant classification ندارد
رگرسیون خطی	می‌تواند، بدون لایه مخفی، تابع فعالیت خطی
لاجیستیک رگرسیون	می‌تواند، بدون لایه مخفی، تابع فعالیت سیگنوید
SVM	نمی‌تواند؛ امکان ایجاد و تنظیم حاشیه ندارد.

سوال سوم-۶ (نمره)

بیشتر می‌شود	کمتر می‌شود	تغییری نمی‌کند	بستگی به شرایط دارد
	✓		

با افزایش عدد k تعداد داده آموزشی افزایش و در نتیجه قابلیت تعمیم‌پذیری بیشتر، و خطای اعتبارسنجی کم می‌شود.

سوال چهارم-۱۸ (نمره)

$E_{Valid} \gg E_{Train}$	عارضه: قابلیت مدل خیلی بالا است، تعداد داده آموزش خیلی کم است. راه حل: انواع روش رگولاریزیشن، کاهش پیچیدگی مدل، افزایش داده آموزشی
$E_{Valid} \cong E_{Train}$	عارضه: پیچیدگی مدل پایین است و مقدار داده در حد پیچیدگی مدل کم است. راه حل: افزایش پیچیدگی مدل، افزایش زمان آموزش برای بررسی کاهش احتمالی خطای آموزش.

سوال پنجم-۳۰ (نمره)  $R_K = 1 + \sum_{f=1}^K (F_f - 1) \prod_{l=0}^{f-1} S_l$

سوال ششم - ۱۶ (نمره)

الف	ب	ج	د	ه
IN	GN	LN	GhostNorm (در کلاس مطرح نشده)	BN

برای هریک باید دلایل (شیوه محاسبه متوسط و واریانس روی مجموعه داده گفته شود)

سوال هفتم -۵ (نمره). تنظیم تعداد کانال‌ها به عدد دلخواه (کاهش و یا افزایش)

سوال هشتم - ۵ (نمره) بله، حتماً باید ذخیره شود، در غیراین صورت آدرس مسیر به روزرسانی ضرایب قبل از لایه از میان می‌رود.

سوال نهم- ۵ (نمره) گزینه اول دامنه بردارگرادیان را کنترل می‌کند، گزینه آخر سطح بزرگی خروجی‌ها را تنظیم می‌کند.

استفاده از Gradient Clipping	استفاده از SkipConnection	استفاده از ReLU بجای سگموئید	لایه Batch Normalization
✓	x	x	✓

مقداردهی مجدد با همان توزیع	مقداردهی مجدد با $Uni(-1,1)$	استفاده از $ReLU$ بجای سگموئید	استفاده از $tanh$ بجای سگموئید
×	✓	✓	×

مشکی از بزرگی ضرایب است که باعث اشباع می شود، گزینه دوم راه حل ارجح است و گزینه سوم، جلوی اشباع را می گیرد.

سوال یازدهم - ۳۰ نمره

$$J^* = \frac{1}{N} \sum_{i=1}^N (y_i - W^T(X_i + \delta_i))^2 = \frac{1}{N} \sum_{i=1}^N ((y_i - W^T X_i) - W^T \delta_i)^2$$

$$J^* = \frac{1}{N} \sum_{i=1}^N (y_i - W^T X_i)^2 - \frac{2}{N} \sum_{i=1}^N ((y_i - W^T X_i) W^T \delta_i) + \frac{1}{N} \sum_{i=1}^N (W^T \delta_i)^2$$

$$J^* = \frac{1}{N} \sum_{i=1}^N (y_i - W^T X_i)^2 - \frac{2}{N} \sum_{i=1}^N ((y_i - W^T X_i) W^T \delta_i) + \frac{1}{N} \sum_{i=1}^N W^T \delta_i \delta_i^T W$$

$$J^* = J - \frac{2}{N} \sum_{i=1}^N ((y_i - W^T X_i) W^T \delta_i) + \frac{1}{N} \sum_{i=1}^N W^T \delta_i \delta_i^T W$$

$$\mathbb{E}\{J^*\} = \mathbb{E}\{J\} - \frac{2}{N} \sum_{i=1}^N ((y_i - W^T X_i) W^T \mathbb{E}\{\delta_i\}) + \frac{1}{N} \sum_{i=1}^N W^T \mathbb{E}\{\delta_i \delta_i^T\} W$$

$$\mathbb{E}\{J^*\} = J + \frac{1}{N} \sum_{i=1}^N W^T \sigma^2 I W = J + \frac{\sigma^2}{N} \sum_{i=1}^N \|W\|_2^2 = J + \sigma^2 \|W\|_2^2$$

صحت روابط جبرخطی، و ... در اخذ نمره کامل اهمیت داشته و در صورت اشتباه نمره بین ۵ الی ۱۰ کسر شده است.

## سوال دوازدهم – ۳۰ نمره

الف) ۵ نمره

$$L = (y - XW)^T(y - XW) + \lambda W^T W = y^T y - 2y^T XW + W^T X^T XW + \lambda W^T W$$

$$\frac{\partial L}{\partial W} = -2X^T y + 2X^T XW + 2\lambda W = 0 \Rightarrow W = (X^T X + \lambda I)^{-1} X^T y$$

ب) ۵ نمره

$$W = (X^T X + \lambda I)^{-1} X^T y = (V \Sigma U^T U \Sigma V^T + \lambda I)^{-1} V \Sigma U^T y = \dots = V(\Sigma^2 + \lambda I)^{-1} \Sigma U^T y = V(\Sigma(I + \lambda \Sigma^{-2}))^{-1} U^T y$$

اثر ضریب هموارسازی:  $diag\left(\frac{1}{\sigma_i}\right) \Rightarrow diag\left(\frac{1}{\sigma_i} \frac{\sigma_i^2}{\sigma_i^2 + \lambda}\right)$  است، اثر مقادیر ویژه بزرگ تغییری نمی‌کند، و اثر مقادیر ویژه کوچک، کاهش می‌یابد

پ) ۵ نمره

$$W(k+1) = W(k) - \eta(-2X^T y + 2X^T XW(k) + 2\lambda W(k)) = W(k) - (2\eta X^T XW(k) + 2\eta \lambda W(k)) + 2\eta X^T y$$

$$W(k+1) = ((1 - 2\eta\lambda)I - 2\eta X^T X)W(k) + 2\eta X^T y = W(k+1) = ((1 - 2\eta\lambda)I - 2\eta V \Sigma U^T U \Sigma V^T)W(k) + 2\eta X^T y$$

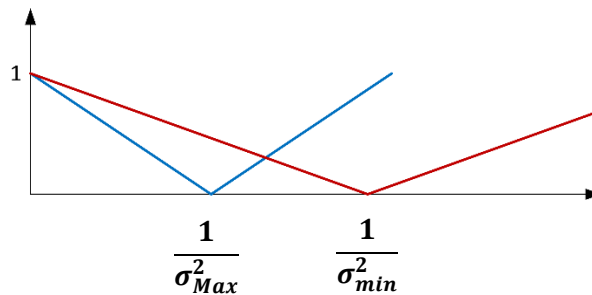
$$W(k+1) = V((1 - 2\eta\lambda)I - 2\eta \Sigma^2)V^T W(k) + 2\eta X^T y$$

ت) ۱۵ نمره

نرخ همگرایی در هر راستا متناسب نمایی با  $|1 - 2\eta\sigma_i^2|$  است، لذا برای حالت بهینه باید مسئله زیر حل شود:

$$\min_{\eta} \max_i |1 - 2\eta\sigma_i^2|$$

با توجه به شکل زیر مقدار بهینه از رابطه زیر بدست می‌آید:  $1 - 2\eta\sigma_{min}^2 = -(1 - 2\eta\sigma_{max}^2)$  که در نتیجه می‌شود:



$$\eta_{opt} = \frac{1}{\sigma_{min}^2 + \sigma_{max}^2}$$