Deep Learning for Visual Recognition
Assignment 1: Machine Learning Basics

Bilge Ulusay
Ewald Bindereif
Ulvi Shukurzade
Cüneyt Erem

November 23, 2020

1 Theoretical Exercises

1.1 Bias of an estimator

$$\hat{\sigma}_m^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \hat{\mu}_m)^2$$

$$\text{Bias}(\hat{\sigma}_m^2) = \mathbb{E}\left[\hat{\sigma}_m^2\right] - \sigma^2 = \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}(x_i - \hat{\mu}_m)^2\right] - \sigma^2 \overset{(1)}{=} \frac{(m-1)\cdot\sigma^2}{m} - \sigma^2 = \frac{-\sigma^2}{m} \overset{(2)}{\neq} 0$$

Step (1) holds due to the hint on the problem sheet and (2) since $\sigma > 0$. It follows that $\hat{\sigma}_m^2$ is a **biased** estimator.

$$\tilde{\sigma}_m^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \hat{\mu}_m)^2$$

$$\text{Bias}(\tilde{\sigma}_m^2) = \mathbb{E}\left[\tilde{\sigma}_m^2\right] - \sigma^2 = \mathbb{E}\left[\frac{1}{m-1}\sum_{i=1}^{m}(x_i - \hat{\mu}_m)^2\right] - \sigma^2 \overset{(1)}{=} \frac{(m-1)\cdot\sigma^2}{m-1} - \sigma^2 = 0$$

It follows that $\tilde{\sigma}_m^2$ is an **unbiased** estimator. Points 5/5

1.2 Bias Variance Trade-off

$$\begin{aligned}\text{Bias}(\hat{\theta})^2 &= (\mathbb{E}[\hat{\theta}] - \theta)^2 = \mathbb{E}[\hat{\theta}]^2 - 2\mathbb{E}[\hat{\theta}]\theta + \theta^2 \\ \text{Var}(\hat{\theta}) &= \mathbb{E}[\hat{\theta}^2] - \mathbb{E}[\hat{\theta}]^2 \\ \text{Bias}(\hat{\theta})^2 + \text{Var}(\hat{\theta}) &= \mathbb{E}[\hat{\theta}^2] - 2\mathbb{E}[\hat{\theta}]\theta + \theta^2\end{aligned}$$

$$\text{MSE}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2] = \mathbb{E}[\hat{\theta}^2 - 2\hat{\theta}\theta + \theta^2] = \mathbb{E}[\hat{\theta}^2] - 2\mathbb{E}[\hat{\theta}]\theta + \theta^2 = \text{Bias}(\hat{\theta})^2 + \text{Var}(\hat{\theta})$$

A nice solution, fewer boring computation steps compared to mine. Points 5/5

1.3 MAP for a conditional likelihood

$$\begin{aligned}
\theta_{\text{MAP}} &= \operatorname*{argmax}_{\theta} p(\theta|X,Y) \\
&\overset{(1)}{=} \operatorname*{argmax}_{\theta} p(X,Y|\theta)p(\theta) \\
&= \operatorname*{argmax}_{\theta} p(Y|X,\theta)p(X|\theta)p(\theta) \\
&\overset{(2)}{=} \operatorname*{argmax}_{\theta} p(Y|X,\theta)p(X)p(\theta) \\
&\overset{(3)}{=} \operatorname*{argmax}_{\theta} p(Y|X,\theta)p(\theta) \\
&\overset{(4)}{=} \operatorname*{argmax}_{\theta} \big(\log p(Y|X,\theta) + \log p(\theta)\big) \\
&= \operatorname*{argmax}_{\theta} \left(\sum_{i=1}^{m} \log p(y^{(i)}|x^{(i)},\theta) + \log p(\theta)\right)
\end{aligned}$$

(1) holds by Bayes' rule: $p(\theta|X,Y) = \frac{p(X,Y|\theta)}{p(X,Y)}$; thus the denominator can be neglected for argmax concerning θ.
(2) holds as θ and X are independent.
(3) holds as $p(X)$ represents a constant and thus can be neglected for argmax concerning θ.
(4) holds after applying the logarithm to the product. ✓ Points 5/5

1.4 Derivatives and the chain rule

We derive the following relationship between the sigmoid function σ and its derivative for $t \in \mathbb{R}$:

$$\begin{aligned}
\sigma'(t) = [(1+e^{-t})^{-1}]' &= -(1+e^{-t})^{-2} \cdot (-e^{-t}) \\
&= \frac{e^{-t}+1-1}{(1+e^{-t})^2} \\
&= \frac{1+e^{-t}}{(1+e^{-t})^2} - \frac{1}{1+e^{-t}} \cdot \frac{1}{1+e^{-t}} \\
&= \frac{1}{1+e^{-t}} \cdot \left(1 - \frac{1}{1+e^{-t}}\right) \\
&= \sigma(t)(1-\sigma(t)) \quad ✓
\end{aligned}$$

Now we calculate the partial derivative of $L(W,b)$ with respect to W_j:

$$\begin{aligned}
\frac{\partial L(W,b)}{\partial W_j} &= \frac{\partial}{\partial W_j}\left(\sum_{i=1}^{N}(y_i - f_{W,b}(x_i))^2\right) \\
&= \sum_{i=1}^{N} \frac{\partial}{\partial W_j}(y_i - f_{W,b}(x_i))^2 \\
&= \sum_{i=1}^{N} 2(y_i - f_{W,b}(x_i))\frac{\partial(y_i - f_{W,b}(x_i))}{\partial W_j} \quad \text{(by chain rule)} \\
&= -2\sum_{i=1}^{N}(y_i - f_{W,b}(x_i))\frac{\partial f_{W,b}(x_i)}{\partial W_j} \\
&= -2\sum_{i=1}^{N}(y_i - f_{W,b}(x_i)) \cdot f_{W,b}(x_i) \cdot (1 - f_{W,b}(x_i)) \cdot x_{ij}, \quad ✓
\end{aligned}$$

where x_{ij} is the j-th component of x_i. The last step follows from the derivation above, because $f_{W,b}$ is a sigmoid function, and the chain rule ($\frac{\partial}{\partial W_j}(\langle W, x_i\rangle + b) = \frac{\partial}{\partial W_j}(W_1 x_{i1} + \cdots + W_d x_{id} + b) = x_{ij}$). ✓

Now we calculate the partial derivative of $L(W, b)$ with respect to b:

$$\begin{aligned}
\frac{\partial L(W,b)}{\partial b} &= \frac{\partial}{\partial b}\left(\sum_{i=1}^{N}(y_i - f_{W,b}(x_i))^2\right) \\
&= \sum_{i=1}^{N}\frac{\partial}{\partial b}(y_i - f_{W,b}(x_i))^2 \\
&= \sum_{i=1}^{N} 2(y_i - f_{W,b}(x_i))\frac{\partial(y_i - f_{W,b}(x_i))}{\partial b} \quad \text{(by chain rule)} \\
&= -2\sum_{i=1}^{N}(y_i - f_{W,b}(x_i))\frac{\partial f_{W,b}(x_i)}{\partial b} \\
&= -2\sum_{i=1}^{N}(y_i - f_{W,b}(x_i)) \cdot f_{W,b}(x_i) \cdot (1 - f_{W,b}(x_i)),
\end{aligned}$$

The last step follows from the derivation above, because $f_{W,b}$ is a sigmoid function, and the chain rule ($\frac{\partial}{\partial b}(\langle W, x_i\rangle + b) = \frac{\partial}{\partial b}(W_1 x_{i1} + \cdots + W_d x_{id} + b) = 1$).

The gradient of $L(W, b)$ with respect to the vector W is equal to the vector, which contains all the partial derivatives of $L(W, b)$ with respect to a single component of W:

$$\nabla_W L(W,b) = \left(\frac{\partial L(W,b)}{\partial W_1}, \ldots, \frac{\partial L(W,b)}{\partial W_d}\right)^T$$

We can update W_j by the equation

$$W_j = W_j - \eta\frac{\partial L(W,b)}{\partial W_j}, \tag{1}$$

and b by

$$b = b - \eta\frac{\partial L(W,b)}{\partial b}, \tag{2}$$

where η denotes the learning rate. Points 5/5

Theory - Points 20/20