Exercise 1 - Variance

Show that for two independent random variables, X, Y and arbitraty $a, b \in \mathbb{R}$, the following equality holds

$$\mathbf{Var}(aX + bY) = a^2 \cdot \mathbf{Var}(X) + b^2 \cdot \mathbf{Var}(Y).$$

Solution

First, we use the definition of variance and rewrite the left hand side as

$$\mathbf{Var}(aX + bY) = \mathbb{E}\left[(aX + bY)^2\right] - \mathbb{E}[aX + bY]^2.$$

Next, we expand the squares for each of the terms on the right hand side:

$$\begin{aligned}
\mathbb{E}\left[(aX + bY)^2\right] &= \mathbb{E}\left[a^2X^2 + 2abXY + b^2Y^2\right] \\
&= a^2\mathbb{E}\left[X^2\right] + 2ab\mathbb{E}[XY] + b^2\mathbb{E}\left[Y^2\right] \\
&= a^2\mathbb{E}\left[X^2\right] + 2ab\mathbb{E}[X]\,\mathbb{E}[Y] + b^2\mathbb{E}\left[Y^2\right], \\
\mathbb{E}[aX + bY]^2 &= \left(a\mathbb{E}[X] + b\mathbb{E}[Y]\right)^2 \\
&= a^2\mathbb{E}[X]^2 + 2ab\mathbb{E}[X]\,\mathbb{E}[Y] + b^2\mathbb{E}[Y]^2.
\end{aligned}$$

Subtracting the two terms, we get

$$\begin{aligned}
&\mathbb{E}\left[(aX + bY)^2\right] - \mathbb{E}[aX + bY]^2 \\
&\quad = a^2\mathbb{E}\left[X^2\right] + 2ab\mathbb{E}[X]\,\mathbb{E}[Y] + b^2\mathbb{E}\left[Y^2\right] - a^2\mathbb{E}[X]^2 - 2ab\mathbb{E}[X]\,\mathbb{E}[Y] - b^2\mathbb{E}[Y]^2 \\
&\quad = a^2\mathbb{E}\left[X^2\right] - a^2\mathbb{E}[X]^2 + b^2\mathbb{E}\left[Y^2\right] - b^2\mathbb{E}[Y]^2 \\
&\quad = a^2(\mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2) + b^2\left(\mathbb{E}\left[Y^2\right] - \mathbb{E}[Y]^2\right) \\
&\quad = a^2 \cdot \mathbf{Var}(X) + b^2 \cdot \mathbf{Var}(Y).
\end{aligned}$$

Exercise 2 - Variance / Bias Decomposistion

Let $D = \{(x_i, y_i) | i = 1 \ldots n\}$ be a dataset obtained from the true underlying data distribution P, i.e. $D \sim P^n$. And let $h_D(\cdot)$ be a classifier trained on D. Show the variance bias decomposition

$$\underbrace{\mathbb{E}_{D,x,y}\left[(h_D(x) - y)^2\right]}_{\text{Expected test error}} = \underbrace{\mathbb{E}_{D,x}\left[(h_D(x) - \hat{h}(x))^2\right]}_{\text{Variance}} + \underbrace{\mathbb{E}_{x,y}\left[(\hat{y}(x) - y)^2\right]}_{\text{Noise}} + \underbrace{\mathbb{E}_{x}\left[(\hat{h}(x) - \hat{y}(x))^2\right]}_{\text{Bias}^2}$$

where $\hat{h}(x) = \mathbb{E}_{D \sim P^n}[h_D(x)]$ is the expected regressor over possible training sets, given the learning algorithm $\mathcal{A}$ and $\hat{y}(x) = \mathbb{E}_{y|x}[y]$ is the expected label given x. As mentioned in the lecture, labels might not be deterministic given x. To carry out the proof, proceed in the following steps:

(a) Show that the following identity holds

$$\mathbb{E}_{D,x,y}\left[[h_D(x) - y]^2\right] = \mathbb{E}_{D,x}\left[(\hat{h}_D(x) - \hat{h}(x))^2\right] + \mathbb{E}_{x,y}\left[\left(\hat{h}(x) - y\right)^2\right]. \tag{1}$$

(b) Next, show

$$E_{x,y}\left[\left(\hat{h}(x) - y\right)^2\right] = E_{x,y}\left[(\hat{y}(x) - y)^2\right] + E_x\left[\left(\hat{h}(x) - \hat{y}(x)\right)^2\right] \tag{2}$$

which completes the proof by substituting (2) into (1).

Solution

(a) First, we reformulate (1) as

$$\begin{aligned}\mathbb{E}_{D,x,y}\left[[h_D(x)-y]^2\right] &= \mathbb{E}_{D,x,y}\left[\left[\left(h_D(x)-\hat{h}(x)\right)+\left(\hat{h}(x)-y\right)\right]^2\right] \\ &= \mathbb{E}_{x,D}\left[(\hat{h}_D(x)-\hat{h}(x))^2\right]+2\,\mathbb{E}_{x,y,D}\left[\left(h_D(x)-\hat{h}(x)\right)\left(\hat{h}(x)-y\right)\right]+\mathbb{E}_{x,y}\left[\left(\hat{h}(x)-y\right)^2\right]\end{aligned}$$

Next, we note that the second term in the above equation is zero because

$$\begin{aligned}\mathbb{E}_{D,x,y}\left[\left(h_D(x)-\hat{h}(x)\right)\left(\hat{h}(x)-y\right)\right] &= \mathbb{E}_{x,y}\left[\mathbb{E}_D\left[h_D(x)-\hat{h}(x)\right]\left(\hat{h}(x)-y\right)\right] \\ &= \mathbb{E}_{x,y}\left[\left(\mathbb{E}_D\left[h_D(x)\right]-\hat{h}(x)\right)\left(\hat{h}(x)-y\right)\right] \\ &= \mathbb{E}_{x,y}\left[\left(\hat{h}(x)-\hat{h}(x)\right)\left(\hat{h}(x)-y\right)\right] \\ &= \mathbb{E}_{x,y}\left[0\right] \\ &= 0\ .\end{aligned}$$

(b) The proof here, is similar. We start by reformulating the second term in (2) as

$$\begin{aligned}\mathbb{E}_{x,y}\left[\left(\hat{h}(x)-y\right)^2\right] &= \mathbb{E}_{x,y}\left[\left(\hat{h}(x)-\bar{y}(x))+(\bar{y}(x)-y\right)^2\right] \\ &= \mathbb{E}_{x,y}\left[(\hat{y}(x)-y)^2\right]+\mathbb{E}_x\left[\left(\hat{h}(x)-\hat{y}(x)\right)^2\right]+2\,\mathbb{E}_{x,y}\left[\left(\hat{h}(x)-\hat{y}(x)\right)(\hat{y}(x)-y)\right]\end{aligned}$$

Here, the third term is zero which follows from an analogous derivation as in (a). Thus, we have

$$\begin{aligned}\mathbb{E}_{x,y}\left[\left(\hat{h}(x)-\hat{y}(x)\right)(\hat{y}(x)-y)\right] &= \mathbb{E}_x\left[\mathbb{E}_{y|x}\left[\hat{y}(x)-y\right]\left(\hat{h}(x)-\hat{y}(x)\right)\right] \\ &= \mathbb{E}_x\left[\mathbb{E}_{y|x}\left[\hat{y}(x)-y\right]\left(\hat{h}(x)-\hat{y}(x)\right)\right] \\ &= \mathbb{E}_x\left[\left(\hat{y}(x)-\mathbb{E}_{y|x}\left[y\right]\right)\left(\hat{h}(x)-\hat{y}(x)\right)\right] \\ &= \mathbb{E}_x\left[(\hat{y}(x)-\hat{y}(x))\left(\hat{h}(x)-\hat{y}(x)\right)\right] \\ &= \mathbb{E}_x\left[0\right] \\ &= 0\end{aligned}$$

Exercise 3 - Ensembling

Download the file `ex06-ensembling.ipynb` from quercus. It contains basic Pytorch code training a classifier on MNIST. Modify that code such that it trains an ensemble of 5-10 neural networks and computes their average prediction once trained.