

**Exercise 1 - Variance**

Show that for two independent random variables,  $X, Y$  and arbitrary  $a, b \in \mathbb{R}$ , 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

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

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

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

Subtracting the two terms, we get

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

**Exercise 2 - Variance / Bias Decomposition**

Let  $D = \{(x_i, y_i) | i = 1 \dots n\}$  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

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

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

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

(b) Next, show

$$\mathbb{E}_{x,y}[(\hat{h}(x) - y)^2] = \mathbb{E}_{x,y}[(\hat{y}(x) - y)^2] + \mathbb{E}_x[(\hat{h}(x) - \hat{y}(x))^2] \quad (2)$$

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

**Solution**

(a) First, we reformulate (1) as

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

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

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

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

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

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

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

**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.