Assignment 3

How to Split a Sample into Training and Test Sets (20 points)

Assume that we have a fixed split of our sample S of i.i.d variables, into S^{Train} and S^{Test}, with $|S^{\text{Test}}| = n^{\text{Test}}$. We train $\hat{h}^*_{S^{\text{Train}}}$ on S^{Train}. We have that since the data in S is i.i.d. the test loss $\hat{L}(\hat{h}^*_{S^{\text{Train}}}, S^{\text{Test}})$ is an unbiased estimator of the true loss $L(\hat{h}^*_{S^{\text{Train}}})$. We assume that the loss function is bounded between $[0,1]$. Now applying Hoeffdings inequality, we have that given $\delta \in (0,1)$

$$\mathbb{P}\left(L(\hat{h}^*_{S^{\text{Train}}}) - \hat{L}(\hat{h}^*_{S^{\text{Train}}}, S^{\text{Test}}) \geq \sqrt{\frac{\log 1/\delta}{2n^{\text{Test}}}}\right) \leq \delta$$

Equivalently

$$\mathbb{P}\left(L(\hat{h}^*_{S^{\text{Train}}}) \leq \hat{L}(\hat{h}^*_{S^{\text{Train}}}) + \sqrt{\frac{\log 1/\delta}{2n^{\text{Test}}}}\right) \geq 1 - \delta$$

We have that the following generalisation bound holds with probability at least $1 - \delta$

$$L(\hat{h}^*_{S^{\text{Train}}}) \leq \hat{L}(\hat{h}^*_{S^{\text{Train}}}) + \sqrt{\frac{\ln(1/\delta)}{2n^{\text{Test}}}}$$

Confidence intervals for Bernoulli distributions (15 points)

Looking at the last formula at page 14 of Seldin's lecture notes, which we can use as we are dealing with Bernoulli random variables (and denoting $\bar{X} = 1/n \sum_{i=1}^{n} X_i$), we have that

$$\mathbb{P}\left(|\bar{X} - \mu| \geq \sqrt{\frac{\ln 2/\delta}{2n}}\right) = \mathbb{P}\left(|\mu - \bar{X}| \geq \sqrt{\frac{\ln 2/\delta}{2n}}\right) \leq \delta$$

And thus

$$\mathbb{P}\left(|\mu - \bar{X}| \leq \sqrt{\frac{\ln 2/\delta}{2n}}\right) \geq 1 - \delta$$

Further we have that

$$\mathbb{P}\left(|\mu - \bar{X}| \leq \sqrt{\frac{\ln 2/\delta}{2n}}\right) = \mathbb{P}\left(-\sqrt{\frac{\ln 2/\delta}{2n}} \leq \mu - \bar{X} \leq \sqrt{\frac{\ln 2/\delta}{2n}}\right) = \mathbb{P}\left(\bar{X} - \sqrt{\frac{\ln 2/\delta}{2n}} \leq \mu \leq \sqrt{\frac{\ln 2/\delta}{2n}} + \bar{X}\right)$$

Setting $\delta = 0.03$ and our sample S, we get the following two-sided 0.97-CI for μ,

$$\mathbb{P}(\mu \in [0.5594615, 0.68204118]) \geq 0.97$$

3 Big and Small (25 points)

Let h_A and h_B be two hypotheses for a classification problem. Assume that h_A was trained on S_A with $|S_A| = n_A = 9000$, and that h_B was trained on S_B with $|S_B| = n_B = 1000$. Assume that all data is i.i.d. and that we are using ℓ bounded in the $[0, 1]$ interval. We test h_A on S_b and vice versa, and get that $\hat{L}(h_A, S_b) = 0.03$ and that $\hat{L}(h_B, S_A) = 0.06$

3.1

We need to pick a classifier. We would prefer the classifier that minimizes expected loss. Recall that the test losses of h_A on S_B and h_B on S_A are unbiased estimates of the expected loss. We see that h_A is superior to h_B in regards to their unbiased point estimates of their expected loss, with only half the test loss of h_B. This speaks in favor of using h_A. However, h_A is tested on only a ninth of the samples compared to h_B. Therefore we would expect the test loss of h_B to more representative of the expected loss of h_B, compared to more uncertainty for h_A. That is, we could be in a situation, where h_A performs very well on the test set, but the test loss is not representative of the true loss. This speaks in favor of using h_B, as we have a better estimate of the loss of h_B.

We can formalis the above by providing generalization bounds for the hypotheses. We have a hypothesis set $\mathcal{H} = \{h_A, h_B\}$. By selecting h out of this set, we introduce bias. By using Hoeffdings inequality before selection we have that, individually

$$\mathbb{P}\left(L(h_A) - \hat{L}(h_A, S_B) \geq \sqrt{\frac{\log 2/\delta}{2n_B}}\right) \leq \frac{\delta}{2} \quad \text{and} \quad \mathbb{P}\left(L(h_B) - \hat{L}(h_B, S_A) \geq \sqrt{\frac{\log 2/\delta}{2n_A}}\right) \leq \frac{\delta}{2}$$

We mimick the proof of theorem 3.2 to arrive at a generalisation bound. We introduce the notation that S_{test} is the sample complementary to the hypothesis, that is if $h = h_A$ then $S_{\text{test}} = S_B$, with $S_{\text{test}} = n_{\text{test}}$.

$$\mathbb{P}\left(\exists h \in \mathcal{H} : L(h) \geq \hat{L}(h, S_{\text{test}}) + \sqrt{\frac{\log 2/\delta}{2n_{\text{test}}}}\right) = \mathbb{P}\left(\left(L(h_A) - \hat{L}(h_A, S_B) \geq \sqrt{\frac{\log 2/\delta}{2n_B}}\right) \cup \left(L(h_B) - \hat{L}(h_B, S_A) \geq \sqrt{\frac{\log 2/\delta}{2n_A}}\right)\right)$$

Now applying the union bound, (for a measure $\mu(\cdot)$ we have that $\mu(A \cup B) \leq \mu(A) + \mu(B)$), we have that

$$\mathbb{P}\left(\exists h \in \mathcal{H} : L(h) \geq \hat{L}(h, S_{\text{test}}) + \sqrt{\frac{\log 2/\delta}{2n_{\text{test}}}}\right) \leq \mathbb{P}\left(L(h_A) - \hat{L}(h_A, S_B) \geq \sqrt{\frac{\log 2/\delta}{2n_B}}\right) + \mathbb{P}\left(L(h_B) - \hat{L}(h_B, S_A) \geq \sqrt{\frac{\log 2/\delta}{2n_A}}\right)$$

Now using Hoeffdings inequality, we have that

$$\mathbb{P}\left(\exists h \in \mathcal{H} : L(h) \geq \hat{L}(h, S_{\text{test}}) + \sqrt{\frac{\log 2/\delta}{2n_{\text{test}}}}\right) \leq \frac{\delta}{2} + \frac{\delta}{2} = \delta$$

Alternatively it holds for all hypoteses $h \in \mathcal{H}$, that

$$L(h) \leq \hat{L}(h, S_{\text{test}}) + \sqrt{\frac{\log 2/\delta}{2n_{\text{test}}}}$$

With probability at least $1 - \delta$ for $\delta \in (0, 1)$. For we can thus conclude, that jointly we have the generalisation bounds

$$L(h_A) \leq \hat{L}(h_A, S_B) + \sqrt{\frac{\log 2/\delta}{2n_B}}$$
$$L(h_B) \leq \hat{L}(h_B, S_A) + \sqrt{\frac{\log 2/\delta}{2n_A}}$$

With probability $1 - \delta$. Now, plugging in $\delta = 0.05$ and the remaining variables, we have that

$$L(h_A) \leq 0.03 + 0.0429 = 0.0729$$
$$L(h_B) \leq 0.06 + 0.0143 = 0.0743$$

With at least 95% probability. We therefore pick h_A with the generalisation bound $L(h_A) \leq 0.0729$ with at least 95% probability.

3.2

Analogous to part one, this time plugging in $\delta = 0.01$, we have that

$$L(h_A) \leq 0.03 + 0.0514 = 0.0814$$
$$L(h_B) \leq 0.06 + 0.0172 = 0.0772$$

With at least 99% probability. We therefore pick h_B in this case with the generalisation bound $L(h_B) \leq 0.0772$ with at least 99/% probability.

4 Preprocessing (20 points)

4.1 9.1 (4 points)

Let d denote the euclidean distance metric. Then

$$d(\text{Mr.Good}, \text{Mr.Unknown}) = \sqrt{(47-21)^2 + (35-36)^2} = 26.019$$
$$d(\text{Mr.Bad}, \text{Mr.Unknown}) = \sqrt{(22-21)^2 + (40-36)^2} = \sqrt{10}$$

Mr. Unknown's nearest neighbor is therefore Mr. Bad, and he should therefore not be given credit according to the nearest neighbor algorithm. Now consider the case where we are measuring income in dollars, instead of thousands of dollars. We then have that

$$d(\text{Mr.Good}, \text{Mr.Unknown}) = \sqrt{(47-21)^2 + (35000-36000)^2} = 1000.337943$$
$$d(\text{Mr.Bad}, \text{Mr.Unknown}) = \sqrt{(22-21)^2 + (40000-36000)^2} = 4000.000126$$

And now, according to the the nearest neigbor algorithm, Mr. Unknown should be given credit.

4.2 9.2 (4 points)

Let

$$X = \begin{pmatrix} --- & x_1^T & --- \\ & \vdots & \\ --- & x_n^T & --- \end{pmatrix}$$

and $\bar{x} = \frac{1}{n}\sum_{i=1} x_i^n$. Define $Z = X - \mathbf{1}\bar{x}^T$, and $\gamma = I - 1/n\mathbf{1}\mathbf{1}^T$. We want to show that $Z = \gamma X$. As matrixmultiplication is distributive,

$$\gamma X = X - \frac{1}{n}\mathbf{1}\mathbf{1}^T X = X - \frac{1}{n}\mathbf{1}\sum_{i=1}^{n} x_i^T = X - \mathbf{1}\frac{1}{n}\sum_{i=1}^{n} x_i^T = X - \mathbf{1}\bar{x}^T$$

Which yields the desired.

4.3 9.4 (12 points)

(a)

Let $\hat{x}_1, \hat{x}_2 \sim \mathcal{N}(0,1)$ be independent. Then the random variable $\hat{x} = (\hat{x}_1, \hat{x}_2)$ follows a 2-dimensional Gaussian distribution with zero mean and identity variance, that is $\hat{x} \sim \mathcal{N}_2(0, I_2)$. We know that for a matrix A, we have that $A\hat{x} \sim \mathcal{N}_2(0, AA^T)$. Now define

$$A = \begin{pmatrix} 1 & 0 \\ \sqrt{1-\epsilon^2} & \epsilon^2 \end{pmatrix}$$

Then

$$AA^T = \begin{pmatrix} 1 & \sqrt{1-\epsilon^2} \\ \sqrt{1-\epsilon^2} & 1 \end{pmatrix}$$

And thereby

$$x = A\hat{x} = \begin{pmatrix} \hat{x}_1 \\ \sqrt{1-\epsilon^2}\hat{x}_1 + \epsilon\hat{x}_2 \end{pmatrix} \sim \mathcal{N}_2\left(0, \begin{pmatrix} 1 & \sqrt{1-\epsilon^2} \\ \sqrt{1-\epsilon^2} & 1 \end{pmatrix}\right)$$

We conclude that $\mathrm{Var}(x_1) = \mathrm{Var}(x_2) = 1$ and $\mathrm{Cov}(x_1, x_2) = \sqrt{1-\epsilon^2}$

(b)

Suppose $f(\hat{x}) = \hat{w}_1\hat{x}_1 + \hat{w}_2\hat{x}_2$. We want to define w_1, w_2, such that $f(x) = w_1x_1 + w_2x_2$. We have that

$$\begin{aligned} f(x) &= w_1x_1 + w_2x_2 \\ &= w_1\hat{x}_1 + w_2(\sqrt{1-\epsilon^2}\hat{x}_1 + \epsilon\hat{x}_2) \\ &= \hat{x}_1(w_1 + w_2\sqrt{1-\epsilon^2}) + w_2\epsilon\hat{x}_2 \end{aligned}$$

Now setting

$$\begin{pmatrix} \hat{w}_1 \\ \hat{w}_2 \end{pmatrix} = \begin{pmatrix} w_1 + w_2\sqrt{1-\epsilon^2} \\ w_2\epsilon \end{pmatrix}$$

yields that

$$f(x) = w_1x_1 + w_2x_2$$

(c)

Consider the target $f(\hat{x}) = \hat{x}_1 + \hat{x}_2$, and the reguralization constraint $w_1 + w_2 \leq C$. We want to implement the target, that is

$$\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} \hat{w}_1 \\ \hat{w}_2 \end{pmatrix} = \begin{pmatrix} w_1 + w_2\sqrt{1-\epsilon^2} \\ w_2\epsilon \end{pmatrix}$$

We clearly have that $w_2 = 1/\epsilon$. Inserting this into the first equation yields

$$1 = w_1 + \frac{\sqrt{1-\epsilon^2}}{\epsilon} \Leftrightarrow w_1 = \frac{\epsilon - \sqrt{1-\epsilon^2}}{\epsilon}$$

Plugging this into our regularization constraint, gives us that

$$C \geq w_1^2 + w_2^2 = \frac{1}{\epsilon^2} + \frac{\epsilon - \sqrt{1-\epsilon^2}}{\epsilon^2} = 2\frac{1-\epsilon\sqrt{1-\epsilon^2}}{\epsilon^2}$$

And the minimum C is thus $C = 2\frac{1-\epsilon\sqrt{1-\epsilon^2}}{\epsilon^2}$.

(d)

As the correlation increases, that is $\epsilon \longrightarrow 0$, we have that

$$2\frac{1-\epsilon\sqrt{1-\epsilon^2}}{\epsilon^2} \longrightarrow \infty$$

And the maximum amount of reguralization goes to infinity.

5 Variable importance (20 points)

Assume that we use one-hot-encoding for generating our model matrix and observe $x_1, \cdots, x_k$. Assuming classes $(C_i)_{i\in\{1,n\}}$, our model matrix with a constant term would look like

$$X = \begin{pmatrix} 1 & \mathbb{1}(x_1 = C_1) & \cdots & \mathbb{1}(x_1 = C_n) \\ 1 & \mathbb{1}(x_2 = C_1) & \cdots & \mathbb{1}(x_2 = C_n) \\ & & \vdots & \\ 1 & \mathbb{1}(x_k = C_1) & \cdots & \mathbb{1}(x_k = C_n) \end{pmatrix}$$

Notice that the columns in X are not linearly independent, as we can write

$$1 = \sum_{i=1}^{n} \mathbb{1}(x_j = C_i)$$

The linear model, is given by the linear subspace

$$L = \{Xw | w \in \mathbb{R}^{n+1}\}$$

Notice the $n+1$-term, which is there as we have used one-hot-encoding on n classes, while we have also augmented the model matrix with the constant 1-vector. Now, as the columns in X are linearly dependent, they do *not* form a basis for L. Therefore the basis expansion of Xw is not unique. That is, when we perform linear regression, we would obtain w^*, but this w^* will not be unique. In particular, we could parametrize the regression solution space (that is, the space of w such that the empirical loss is minimal), using only n linearly independent vectors and a free scalar corresponding to the constraint $1 = \sum_{i=1}^{n} \mathbb{1}(x_j = C_i)$. That is, we would have infinitely many solutions to our regression problem. We would not have this problem, if we had used dummy variables instead, as the columns of X would be linearly independent.

This also explains why it would become meaningless to interpret the variable importance if one-hot encoding was used. If someone claimed something about the variable importance using the optimal weights w^*, another equally valid claim about variable importance could be made using another optimal weight configuration $\bar{w}^*$. Clearly the variable importance indicated by w^* and $\bar{w}^*$ cannot be correct simoultaneously, and we have no way of deciding which is the correct interpreation. Therefore it would be difficult to interpret variable importance if one-hot encoding was used.