ECEN 649 Pattern Recognition – Spring 2016

Problem Set 3

Due on: Apr 18

1. Let $\mathbf{X} \in R^d$ be a feature set of size d. An additional feature $X_0 \in R$ is a redundant ot "noisy" feature if there is no improvement in discrimination upon joining X_0 to $\mathbf{X}$, i.e., if $\varepsilon^*(\mathbf{X}, Y) = \varepsilon^*(\mathbf{X}', Y)$, where $\mathbf{X}' = (\mathbf{X}, X_0) \in R^{d+1}$. Show that a sufficient condition for this undesirable situation is that X_0 be independent of $(\mathbf{X}, Y)$.

   Solution: Using the independence of $(\mathbf{X}, Y)$ from X_0, we have

   $$\eta'(\mathbf{X}') = P(Y = 1 \mid \mathbf{X}') = P(Y = 1 \mid \mathbf{X}, X_0) = P(Y = 1 \mid \mathbf{X}) = \eta(\mathbf{X})$$

   The identity $P(Y = 1 \mid \mathbf{X}, X_0) = P(Y = 1 \mid \mathbf{X})$ can be shown easily by using the definition of conditional probability and the facts that $P(Y, \mathbf{X}, X_0) = P(Y, \mathbf{X})P(X_0)$ and $P(\mathbf{X}, X_0) = P(\mathbf{X})P(X_0)$. In other words, X_0 is independent of Y *given* $\mathbf{X}$.

   It follows that

   $$\varepsilon^*(\mathbf{X}', Y) = E[\min\{\eta'(\mathbf{X}'), 1 - \eta'(\mathbf{X}')\}] = E[\min\{\eta(\mathbf{X}), 1 - \eta(\mathbf{X})\}] = \varepsilon^*(\mathbf{X}, Y)$$

   as required. (The second equality can be checked by writing out the integrals.)

2. You are given that the classification error ε_n and an error estimator $\hat{\varepsilon}_n$ are distributed, as a function of the random sample S_n, as two Gaussians

   $$\varepsilon_n \sim N(\varepsilon^* + 1/n, 1/n^2)\,, \quad \hat{\varepsilon}_n \sim N(\varepsilon^* - 1/n, 1/n^2)\,,$$

   where ε^* is the Bayes error. Find the bias of $\hat{\varepsilon}_n$ as a function of n and plot it; is this estimator optimistically or pessimisitically biased? Find the deviation variance, RMS, tail probabilities, and correlation coefficient as a function of n and plot them. Assume that $\mathrm{Cov}(\varepsilon_n, \hat{\varepsilon}_n) = 1/(2n^2)$.

   Solution: We have $E[\varepsilon_n] = \varepsilon^* + 1/n$ and $E[\hat{\varepsilon}_n] = \varepsilon^* - 1/n$. Therefore

   $$\mathrm{Bias}(\hat{\varepsilon}_n) = E[\hat{\varepsilon}_n] - E[\varepsilon_n] \;=\; -2/n\,,$$

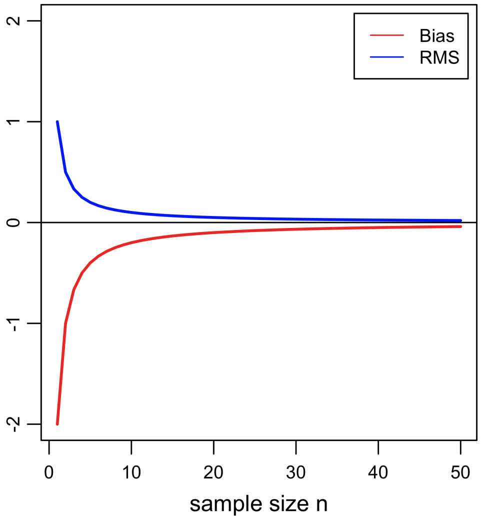
   and the estimator is optimistically biased for all sample sizes. On the other hand, $\mathrm{Var}(\varepsilon_n) = \mathrm{Var}(\hat{\varepsilon}_n) = 1/n^2$, and $\mathrm{Cov}(\varepsilon_n, \hat{\varepsilon}_n) = 1/(2n^2)$, thus

   $$\mathrm{Var}_d(\hat{\varepsilon}_n) = \mathrm{Var}(\hat{\varepsilon}_n - \varepsilon_n) = \mathrm{Var}(\hat{\varepsilon}_n) + \mathrm{Var}(\varepsilon_n) - 2\mathrm{Cov}(\hat{\varepsilon}_n, \varepsilon_n) \;=\; 1/n^2$$

   and hence the RMS is given by

   $$\mathrm{RMS}(\hat{\varepsilon}_n) = \sqrt{E[(\hat{\varepsilon}_n - \varepsilon_n)^2]} = \sqrt{\mathrm{Var}_d(\hat{\varepsilon}_n) + \mathrm{Bias}(\hat{\varepsilon}_n)^2} \;=\; \sqrt{5}/n\,.$$

   The bias and RMS are plotted below.



3. You are given that an error estimator $\hat{\varepsilon}_n$ is related to the classification error ε_n through the simple model

$$\hat{\varepsilon}_n = \varepsilon_n + Z\,,$$

where the conditional distribution of the random variable Z given the training data S_n is Gaussian, $Z \sim N(0, 1/n^2)$. Is $\hat{\varepsilon}_n$ randomized or nonrandomized? Find the internal variance and variance of $\hat{\varepsilon}_n$. What happens as the sample size grows without bound?

Solution: Since $\hat{\varepsilon}_n$ depends on Z, which is random given S_n, the error estimator is randomized. Its internal variance is given by

$$V_{\text{int}} = \text{Var}(\hat{\varepsilon}_n \mid S_n) = \text{Var}(\varepsilon_n + Z \mid S_n) = \text{Var}(Z \mid S_n) = \frac{1}{n^2}\,.$$

The variance can be obtained from the conditional variance formula:

$$\text{Var}(\hat{\varepsilon}_n) = E[V_{\text{int}}] + \text{Var}(E[\hat{\varepsilon}_n \mid S_n]) = \frac{1}{n^2} + \text{Var}(\varepsilon_n + 0) = \frac{1}{n^2} + \text{Var}(\varepsilon_n)\,.$$

As the sample size grows without bound, the internal variance tends to zero, and the variance of the error estimator becomes equal to just the variance of the true classification error (typically, this will converge to zero as well).

4. You would like to use the test-set error estimator in a classification problem.

   (a) If you are given that $\text{Var}(\varepsilon_n) \leq 5 \times 10^{-5}$, find the minimum number of test samples m that will guarantee that the standard deviation of the test-set error estimator $\hat{\varepsilon}_{n,m}$ will be at most 1%.

   Solution: We need $\text{Std}(\hat{\varepsilon}_{n,m}) \leq 10^{-2}$, that is,

$$\text{Var}(\hat{\varepsilon}_{n,m}) = E[V_{\text{int}}] + \text{Var}[\varepsilon_n] \leq 10^{-4}.$$

   This will be guaranteed if $V_{\text{int}} \leq 10^{-4} - 5 \times 10^{-5} = 5 \times 10^{-5}$. But we know that

$$V_{\text{int}} \leq \frac{1}{4m}$$

so we need

$$\frac{1}{4m} \leq 5 \times 10^{-5} \;\Rightarrow\; m \geq \frac{1}{2 \times 10^{-4}} = 5000\,.$$

(b) If you are given that the error of a given classifier is $\varepsilon_n = 0.1$, find the probability that the test-set error estimate $\hat{\varepsilon}_{n,m}$ will be exactly equal to ε_n, if $m = 20$ testing samples are available.

Solution: From the lecture slides, we know that

$$P\left(\hat{\varepsilon}_{n,m} = \frac{k}{m} \,\middle|\, S_n\right) \;=\; \binom{m}{k}\varepsilon_n^k(1-\varepsilon_n)^{m-k}, \quad k = 0,\ldots,m\,.$$

Therefore, with $m = 20$, we have

$$P\left(\hat{\varepsilon}_{n,m} = \varepsilon_n = 0.1 = \frac{2}{20} \,\middle|\, S_n\right) \;=\; \binom{20}{2} 0.1^2\, 0.9^{18} \;\approx\; 0.285\,.$$

5. This problem concerns additional properties of the hold-out estimator.

(a) Show that

$$\mathrm{Var}(\hat{\varepsilon}_{n,m}) \;=\; \frac{E[\varepsilon_n](1-E[\varepsilon_n])}{m} \;+\; \frac{m-1}{m}\,\mathrm{Var}(\varepsilon_n)\,. \tag{1}$$

From this, show that $\mathrm{Var}(\hat{\varepsilon}_{n,m}) \to \mathrm{Var}(\varepsilon_n)$, as the number of testing samples $m \to \infty$.

Solution: From the conditional variance formula, we know that

$$\mathrm{Var}(\hat{\varepsilon}_{n,m}) \;=\; E[V_{\mathrm{int}}] + \mathrm{Var}(E[\hat{\varepsilon}_{n,m}|S_n])$$

But $E[\hat{\varepsilon}_{n,m}|S_n] = \varepsilon_n$ and $V_{\mathrm{int}} = \varepsilon_n(1-\varepsilon_n)/m$, so that

$$\begin{aligned}\mathrm{Var}(\hat{\varepsilon}_{n,m}) \;&=\; E\left[\frac{\varepsilon_n(1-\varepsilon_n)}{m}\right] + \mathrm{Var}(\varepsilon_n) \;=\; \frac{1}{m}\left(E[\varepsilon_n] - E[\varepsilon_n^2]\right) + \mathrm{Var}(\varepsilon_n)\\ &=\; \frac{1}{m}\left(E[\varepsilon_n] - \mathrm{Var}(\varepsilon_n) - E[\varepsilon_n]^2\right) + \mathrm{Var}(\varepsilon_n) \;=\; \frac{E[\varepsilon_n](1-E[\varepsilon_n])}{m} + \frac{m-1}{m}\,\mathrm{Var}(\varepsilon_n).\end{aligned} \tag{2}$$

As $m \to \infty$, the coefficients in this convex combination tend to 0 and 1, respectively, showing that $\mathrm{Var}(\hat{\varepsilon}_{n,m}) \to \mathrm{Var}(\varepsilon_n)$.

(b) Using (1), show that

$$\mathrm{Var}(\varepsilon_n) \;\leq\; \mathrm{Var}(\hat{\varepsilon}_{n,m}) \;\leq\; E[\varepsilon_n](1-E[\varepsilon_n])$$

In particular, this shows that when $E[\varepsilon_n]$ is small, so is $\mathrm{Var}(\hat{\varepsilon}_{n,m})$.

Hint: For any random variable X such that $0 \leq X \leq 1$ with probability 1, one has $\mathrm{Var}(X) \leq E[X](1-E[X])$. (Why?)

Solution: Using the hint, we obtain

$$\mathrm{Var}(\varepsilon_n) \;\leq\; E[\varepsilon_n](1-E[\varepsilon_n]). \tag{3}$$

Applying inequality (3) in (2) yields

$$\mathrm{Var}(\hat{\varepsilon}_{n,m}) \;\leq\; \frac{E[\varepsilon_n](1-E[\varepsilon_n])}{m} + \frac{m-1}{m}\,E[\varepsilon_n](1-E[\varepsilon_n]) \;=\; E[\varepsilon_n](1-E[\varepsilon_n]).$$

But applying inequality (3) in (2) also yields

$$\mathrm{Var}(\hat{\varepsilon}_{n,m}) \;\geq\; \frac{\mathrm{Var}(\varepsilon_n)}{m} + \frac{m-1}{m}\,\mathrm{Var}(\varepsilon_n) \;=\; \mathrm{Var}(\varepsilon_n).$$

(c) Show that the tail probabilities given the training data S_n satisfy:

$$P(|\hat{\varepsilon}_{n,m} - \varepsilon_n| \geq \tau \mid S_n) \;\leq\; 2e^{-2m\tau^2}, \;\text{ for all } \tau > 0\,.$$

Hint: Use Hoeffding's Inequality (DGL Theorem 8.1).

Solution: Let $Z_i = |Y_i - \psi_n(X_i)|$, where (X_i, Y_i) is a test sample, for $i = 1, \ldots, m$. Then the Z_i are independent, bounded random variables, such that Z_i falls into the interval $[0, 1]$ with probability one. In addition, given S_n, we have that $\sum_{i=1}^m Z_i = m\hat{\varepsilon}_{n,m}$, and $E[\sum_{i=1}^m Z_i \,|\, S_n] = mE[Z_1] = m\varepsilon_n$. Therefore, Hoeffding's inequality gives:

$$P(|m\hat{\varepsilon}_{n,m} - m\varepsilon_n| > \theta \,|\, S_n) = P(|\hat{\varepsilon}_{n,m} - \varepsilon_n| > \theta/m \,|\, S_n) \leq 2e^{-2\theta^2/m}, \;\text{ for all } \theta > 0$$

Now let $\tau = \theta/m$, that is, $\theta = m\tau$. This leads to the required inequality:

$$P(|\hat{\varepsilon}_{n,m} - \varepsilon_n| > \tau \,|\, S_n) \leq 2e^{-2m\tau^2}, \;\text{ for all } \tau > 0$$

(d) By using the Strong Law of Large Numbers, show that, given the training data S_n, $\hat{\varepsilon}_{n,m} \to \varepsilon_n$ with probability 1.

Solution: The variables Z_i are independent and identically distributed. They are also bounded, so that all moments are finite. We can thus apply the Strong Law of Large Numbers theorem, which asserts that the sample mean converges to the true mean with probability 1, that is:

$$\frac{1}{m}\sum_{i=1}^m Z_i \to E[Z_1] \;\text{ as } m \to \infty, \;\text{ with probability 1}$$

which is to say that, given S_n,

$$\hat{\varepsilon}_{n,m} \to \varepsilon_n \;\text{ as } m \to \infty, \;\text{ with probability 1}$$

Therefore, as the number of test samples increases to infinity, we have very strong convergence of the hold-out error estimator to the true classification error given S_n.

(e) Repeat item (d), but this time using the result from item (c).

Solution: The same conclusion can be reached by using Hoeffding's inequality (which notably does not require identically-distributed random variables). From part (a), we saw that this implies convergence of $P(|\hat{\varepsilon}_{n,m} - \varepsilon_n| > \tau \,|\, S_n)$ to zero for any $\tau > 0$, as $m \to \infty$ (at an exponential rate); that is, given S_n, $\hat{\varepsilon}_{n,m}$ coverges to ε_n in probability as $m \to \infty$. But, as we showed in the solutions to HW2 (see also DGL Thm A.23), the exponential rate of convergence and the First Borel-Cantelli Lemma transform this convergence in probability to convergence with probability 1.

6. This problem illustrates the very poor (even paradoxical) performance of cross-validation with very small sample sizes. Consider the resubstitution and leave-one-out error estimators $\hat{\varepsilon}_n^r$ and $\hat{\varepsilon}_n^{\mathrm{l}}$ for the 3NN classification rule, with a sample of size $n = 4$ from a mixture of two equally-likely Gaussian populations $\Pi_0 \sim N_d(\boldsymbol{\mu}_0, \Sigma)$ and $\Pi_1 \sim N_d(\boldsymbol{\mu}_1, \Sigma)$. Assume that $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ are far enough apart to make $\delta = \sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T \Sigma^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)} \gg 0$ (in which case the Bayes error is $\varepsilon_{\text{bay}} = \Phi\left(-\delta/2\right) \approx 0$).

   (a) For a sample S_n with $N_0 = N_1 = 2$, which occurs $P(N_0 = 2) = \binom{4}{2} 2^{-4} = 37.5\%$ of the time, show that $\varepsilon_n \approx 0$ but $\hat{\varepsilon}_n^{\mathrm{l}} = 1$.

   Solution: With $n = 4$,computation of the leave-one-out error estimate for the 3NN classification rule proceeds as follows: remove each training point, find the majority label among the remaining three points, and count an error if the majority label is different than the label of the removed point. If $N_0 = N_1 = 2$, it is clear that as each point is removed, the majority label of the remaining points will be opposite to that of the removed point. Therefore, $\hat{\varepsilon}_n^{\mathrm{l}} = 1$. As for the true error, notice that the 3NN decision boundary will correspond to a rough hyperplane located midway between the two class means. Since $\delta \gg 0$, the the class-conditional densities have minimum overlap, and $\varepsilon_n \approx 0$. (Note that $\hat{\varepsilon}_n^{\mathrm{l}}$ would still be 1 even if δ were small.)

   (b) Show that $E[\varepsilon_n] \approx 5/16 = 0.3125$, but $E[\hat{\varepsilon}_n^l] = 0.5$, so that $\text{Bias}(\hat{\varepsilon}_n^l) \approx 3/16 = 0.1875$, and the leave-one-out estimator is far from unbiased.

   Solution: Following the same kind of reasoning as in the previous item, we can construct the following table for each possible configuration of the data

$$
\begin{aligned}
N_0 = 0,\ N_1 = 4 \ &\Rightarrow\ \varepsilon_n \approx \frac{1}{2},\ \hat{\varepsilon}_n^{\mathrm{l}} = 0 \text{ with prob. } P(N_0 = 0) = \binom{4}{0} 2^{-4} = \frac{1}{16} \\
N_0 = 1,\ N_1 = 3 \ &\Rightarrow\ \varepsilon_n \approx \frac{1}{2},\ \hat{\varepsilon}_n^{\mathrm{l}} = \frac{1}{4} \text{ with prob. } P(N_0 = 1) = \binom{4}{1} 2^{-4} = \frac{4}{16} \\
N_0 = 2,\ N_1 = 2 \ &\Rightarrow\ \varepsilon_n \approx 0,\quad \hat{\varepsilon}_n^{\mathrm{l}} = 1 \text{ with prob. } P(N_0 = 2) = \binom{4}{2} 2^{-4} = \frac{6}{16} \\
N_0 = 3,\ N_1 = 1 \ &\Rightarrow\ \varepsilon_n \approx \frac{1}{2},\ \hat{\varepsilon}_n^{\mathrm{l}} = \frac{1}{4} \text{ with prob. } P(N_0 = 3) = \binom{4}{3} 2^{-4} = \frac{4}{16} \\
N_0 = 4,\ N_1 = 0 \ &\Rightarrow\ \varepsilon_n \approx \frac{1}{2},\ \hat{\varepsilon}_n^{\mathrm{l}} = 0 \text{ with prob. } P(N_0 = 4) = \binom{4}{4} 2^{-4} = \frac{1}{16}
\end{aligned}
$$

   Therefore,

$$
\begin{aligned}
E[\varepsilon_n] &\approx \frac{1}{2} \times \frac{10}{16} + 0 \times \frac{6}{16} = \frac{5}{16} \\
E[\hat{\varepsilon}_n^l] &= 0 \times \frac{2}{16} + \frac{1}{4} \times \frac{8}{16} + 1 \times \frac{6}{16} = \frac{1}{2}
\end{aligned}
$$

   (c) Show that $\text{Var}_d(\hat{\varepsilon}_n^l) \approx 103/256 \approx 0.402$, which corresponds to a standard deviation of $\sqrt{0.402} = 0.634$. The leave-one-out estimator is therefore highly-biased and highly-variable in this case.

Solution: Using the table in the solution of part (b),

$$E[(\hat{\varepsilon}_n^{\,l} - \varepsilon_n)^2] \,\approx\, 2 \times \left(\frac{1}{2}\right)^2 \times \frac{1}{16} \,+\, 2 \times \left(\frac{1}{4}\right)^2 \times \frac{4}{16} \,+\, 1^2 \times \frac{6}{16} \,=\, \frac{7}{16}\,.$$

Hence,

$$\mathrm{Var}_d(\hat{\varepsilon}_n^{\,l}) \,=\, E[(\hat{\varepsilon}_n^{\,l} - \varepsilon_n)^2] - E[\hat{\varepsilon}_n^{\,l} - \varepsilon_n]^2 \,\approx\, \frac{7}{16} - \left(\frac{3}{16}\right)^2 \,=\, \frac{103}{256}\,.$$

(d) Show that $\rho(\varepsilon_n, \hat{\varepsilon}_n^{\,l}) \approx -0.98$, i.e., the leave-one-out estimator is almost perfectly negatively correlated with the true classification error.

Solution: Using the table and results in the solution of part (b) and (c), we get

$$\mathrm{Var}(\varepsilon_n) \,=\, E[\varepsilon_n^2] - E[\varepsilon_n]^2 \,\approx\, \left(\frac{1}{2}\right)^2 \times \frac{10}{16} + 0^2 \times \frac{6}{16} - \left(\frac{5}{16}\right)^2 \,=\, \frac{15}{256}$$

$$\mathrm{Var}(\hat{\varepsilon}_n^{\,l}) \,=\, E[(\hat{\varepsilon}_n^{\,l})^2] - E[\hat{\varepsilon}_n^{\,l}]^2 \,=\, 0^2 \times \frac{2}{16} + \left(\frac{1}{4}\right)^2 \times \frac{8}{16} + 1^2 \times \frac{6}{16} - \left(\frac{1}{2}\right)^2 \,=\, \frac{5}{32}$$

$$\mathrm{Cov}(\varepsilon_n, \hat{\varepsilon}_n^{\,l}) \,=\, -\frac{1}{2}(\mathrm{Var}_d(\hat{\varepsilon}_n^{\,l}) - \mathrm{Var}(\varepsilon_n) - \mathrm{Var}(\hat{\varepsilon}_n^{\,l})) \,\approx\, -\frac{1}{2}\left(\frac{103}{256} - \frac{15}{256} - \frac{5}{32}\right) \,=\, -\frac{3}{32}$$

Hence,

$$\rho(\varepsilon_n, \hat{\varepsilon}_n^{\,l}) = \frac{\mathrm{Cov}(\varepsilon_n, \hat{\varepsilon}_n^{\,l})}{\sqrt{\mathrm{Var}(\varepsilon_n)}\sqrt{\mathrm{Var}(\hat{\varepsilon}_n^{\,l})}} \,\approx\, -0.98\,.$$

(e) For comparison, For comparison, show that, although $E[\hat{\varepsilon}_n^{\,r}] = 1/8 = 0.125$, so that $\mathrm{Bias}(\hat{\varepsilon}_n^{\,r}) \approx -3/16 = -0.1875$, which is exactly the negative of the bias of leave-one-out, we have $\mathrm{Var}_d(\hat{\varepsilon}_n^{\,r}) \approx 7/256 \approx 0.027$, for a standard deviation of $\sqrt{7}/16 \approx 0.165$, which is several times smaller than the leave-one-out variance, and $\rho(\varepsilon_n, \hat{\varepsilon}_n^{\,r}) \approx \sqrt{3/5} \approx 0.775$, showing that the resubstitution estimator is highly positively correlated with the true error.

Solution: Computation of the resubstitution error estimate for the 3NN classification rule proceeds as follows: for each training point, find the majority label among the 3 nearest neighbors *including itself* and count an error if the majority label is different than the label of the cirrent point. This yields the table

$$N_0 = 0,\ N_1 = 4 \ \Rightarrow\ \hat{\varepsilon}_n^{\,\mathrm{r}} = 0 \text{ with prob. } P(N_0 = 0) = \binom{4}{0}2^{-4} = \frac{1}{16}$$

$$N_0 = 1,\ N_1 = 3 \ \Rightarrow\ \hat{\varepsilon}_n^{\,\mathrm{r}} = \frac{1}{4} \text{ with prob. } P(N_0 = 1) = \binom{4}{1}2^{-4} = \frac{4}{16}$$

$$N_0 = 2,\ N_1 = 2 \ \Rightarrow\ \hat{\varepsilon}_n^{\,\mathrm{r}} = 0 \text{ with prob. } P(N_0 = 2) = \binom{4}{2}2^{-4} = \frac{6}{16}$$

$$N_0 = 3,\ N_1 = 1 \ \Rightarrow\ \hat{\varepsilon}_n^{\,\mathrm{r}} = \frac{1}{4} \text{ with prob. } P(N_0 = 3) = \binom{4}{3}2^{-4} = \frac{4}{16}$$

$$N_0 = 4,\ N_1 = 0 \ \Rightarrow\ \hat{\varepsilon}_n^{\,\mathrm{r}} = 0 \text{ with prob. } P(N_0 = 4) = \binom{4}{4}2^{-4} = \frac{1}{16}$$

The desired quantities can then be computed exactly as in the previous case of the leave-one-out error estimator.