EE 660 **Homework 6 (Week 11)** Posted: 11/1/2021

Jenkins Due: Mon., 11/8/2021, 11:59 PM PST

1. In class we derived the generalization-error bound for a C-class problem with $C > 2$, from the training-set error, based on the growth function $m_{\mathcal{H}}(2N)$. In this problem, you will derive the generalization-error bound for a C-class problem from the test-set error and from a validation-set error with finite M.

   Throughout this problem:

   let $\underline{\underline{\tilde{C}}}_{\mathcal{D}}$ denote the (in-sample) unnormalized confusion matrix based on dataset $\mathcal{D}$, so that entry $\left(\underline{\underline{\tilde{C}}}_{\mathcal{D}}\right)_{ij}$ = [number of data points labelled $y = j$ that were misclassified as $h = i$];

   also, let $\left(\underline{\underline{C}}_{\text{out}}\right)_{ij} = P\,[(h = i)\, AND\, (y = j)]$ be the ij^{th} entry of the out-of-sample confusion matrix $\underline{\underline{C}}_{\text{out}}$.

   (a) For a given single hypothesis h for the C-class problem (so $h \in \{1,2,\cdots,C\}$) tested using dataset $\mathcal{D}$ that has N data points, give an expression for the total number of points that were misclassified n_{mis} , in terms of the entries $\left(\underline{\underline{\tilde{C}}}_{\mathcal{D}}\right)_{ij}$.

   Also give an expression for the error rate on $\mathcal{D}$, $E_{\mathcal{D}}(h)$, in terms of the entries $\left(\underline{\underline{\tilde{C}}}_{\mathcal{D}}\right)_{ij}$.

   For the out-of-sample confusion matrix, give an expression for the total probability of error $P\,(h \neq y)$ in terms of the entries of $\underline{\underline{C}}_{\text{out}}$. **Hint:** are the events for $\left(\underline{\underline{C}}_{\text{out}}\right)_{ij}$ and $\left(\underline{\underline{C}}_{\text{out}}\right)_{kl}$ mutually exclusive?

   Use these results to give expressions for $\mu = P\,[\text{incorrect classification}]$ and $\nu =$ percent misclassified by h on $\mathcal{D}$.

   Apply Hoeffding Inequality to μ and ν.

   Write the resulting expression in terms of $E_{\mathcal{D}}$ and E_{out} .

   Reformulate to give an expression in the following form:

   $$P[E_{\text{out}}(h) \leq E_{\mathcal{D}}(h) + B(\delta)] \geq 1 - \delta.$$

   in which you fill in for $B(\delta)$. **Hint:** this is similar to what we did in Lecture 7 for the $C = 2$ case.

   Is this a generalization-error bound for test-set error, for a $C > 2$ class problem?

   Comment: As you may have observed in the Midterm Assignment Pr. 1, the generalization-error bound based on a test set can be much tighter than the bound based on a training set and its VC dimension.

(b) Extend the result of (a) to a validation-set error on $\mathcal{D}_{\text{val}}$, in which the hypothesis set has $|\mathcal{H}| = M,\ 0 < M < \infty$.

Hint: does the same technique applying a union bound that we did for the 2-class problem (Lecture 7) apply?

2. This problem concerns the generalization error bound in a transfer learning problem, as given in Lecture 13 (v2.1), Eq. (6).

In this problem you will study the effects of varying N_S, N_T, and α on the cross-domain generalization error bound.

Throughout this problem, let $\varepsilon_{\alpha\beta}$ be everything in the cross-domain generalization-error bound (RHS of Lecture 13 (v2.1) Eq. (6)), except omitting $e^*_{S,T}$. Note that $e^*_{S,T}$ is a constant of the parameters we will be varying.

Also throughout this problem, use the values $d_{VC} = 10$, $\delta = 0.1$, $d_{\mathcal{H}\Delta\mathcal{H}} = 0.1$. However, leave them as variables until you are ready to plot, or until you are asked for a number.

(a) Give the simplified number (to two decimal digits) for $\varepsilon_{\alpha\beta}$, for the following cases:

(i) $N_T = 1,\ N_S = 100,\ \alpha = 0.1,\ 0.5,\ 0.9$

(ii) $N_T = 10,\ N_S = 1000,\ \alpha = 0.1,\ 0.5,\ 0.9$

(iii) $N_T = 100,\ N_S = 10000,\ \alpha = 0.1,\ 0.5,\ 0.9$

(iv) $N_T = 1000,\ N_S = 100000,\ \alpha = 0.1,\ 0.5,\ 0.9$

Tip: put these in a table for easy viewing.

(v) Do any of these sets of numbers assure some degree of generalization (*i.e.*, $\varepsilon_{\alpha\beta} < 0.5$, assuming $e^*_{S,T} \approx 0$)? If so, which?

Comment: As in the supervised learning case, these bounds can be very loose, but evidence indicates the functional dependence of $\varepsilon_{\alpha\beta}$ on its variables still generally apply.

(b) For this part, let $N_S = 1000$ and plot $\varepsilon_{\alpha\beta}$ vs. α for $N_T = 10,\ 100,\ 1000,\ 10000$ (4 curves on one plot), over $0 \le \alpha \le 1$. Answer: what approximate value of α is optimal for each value of N_T ? Try to explain the dependence of $\varepsilon_{\alpha\beta}$ on α for different values of N_T , and any difference in optimal values of α.

(c) For this part, let $N_T = 100$ and plot $\varepsilon_{\alpha\beta}$ vs. α for $N_S = 10,\ 100,\ 1000,\ 10000$ (4 curves on one plot), over $0 \le \alpha \le 1$. Answer: what approximate value of α is optimal for each value of N_T ? Try to explain the dependence of $\varepsilon_{\alpha\beta}$ on α for different values of N_S , and any difference in optimal values of α.

(d) Common default values for α are $\alpha = 0.5$ and $\alpha = \beta$.

   (i) In terms of minimizing the cross-domain generalization-error bound, which default choice looks better (based on your answers to (b) and (c) above)? Is that choice reasonably consistent with your results of (b) and (c)?

   (ii) Give algebraic expressions for $\varepsilon_{\alpha\beta}(\alpha = 0.5)$ and $\varepsilon_{\alpha\beta}(\alpha = \beta)$. Compare them algebraically: can you draw any conclusions about which is lower?

   (iii) Plot $\varepsilon_{\alpha\beta}(\alpha = 0.5)$ vs. N for $\beta = 0.01, 0.1, 0.5$, for $1000 \leq N \leq 100000$ (3 curves on 1 plot). Repeat for $\varepsilon_{\alpha\beta}(\alpha = \beta)$. What conclusions can you draw from the plots?

3. (a) Is it possible to have a covariate shift while satisfying all of:
   $p_S(y|x) = p_T(y|x)$, $p_S(y) = p_T(y)$, $p_S(x|y) = p_T(x|y)$?
   If no, prove your answer; if yes, justify your answer.

   (b) Is it possible to have a covariate shift while satisfying:
   $p_S(y|x) = p_T(y|x)$?
   If no, prove your answer; if yes, justify your answer.

   (c) Is it possible to have a concept shift while satisfying:
   $p_S(y) = p_T(y)$ and $p_S(x) = p_T(x)$?
   If no, prove your answer; if yes, justify your answer.