3. Binary classification, function spaces determined by kernels and regularization

**Assigned reading:** Chapters 4 and 6 of the course notes. Review Section 6.3.

**Optional recommended reading:** Shalev-Shwartz and Ben-David, *Understanding Machine Learning, from Theory to Algorithms*, Chapters 10 & 12.

(Problems and solutions)

1. **[A bit on Rademacher averages.]**
   The proof of the contraction principle in the notes uses the fact if $A, B \subseteq \mathbb{R}^n$ such that $0 \in A$ and $0 \in B$ then $R_n^A(A \cup B) \leq R_n^A(A) + R_n^B(B)$. (Recall that $R_n^A$ has no absolute value.)
   
   (a) Prove the stated property.
   
   **Solution:** Let $a$ and $b$ be defined as functions of $x \in A \cup B$ as follows. If $x \in A$ then $(a, b) = (x, 0)$. If $x \in B \setminus A$ then $(a, b) = (0, x)$. Then for any $x \in A \cup B$ and $\epsilon \in \{-1, 1\}^n$,
   
   $$\frac{1}{n} \sum_{i=1}^{n} x_i \epsilon_i = \frac{1}{n} \sum_{i=1}^{n} a_i \epsilon_i + \frac{1}{n} \sum_{i=1}^{n} b_i \epsilon_i$$  \hspace{1cm} (1)
   
   Taking the supremum over $x$ and average over all $\epsilon \in \{-1, 1\}^n$ yields the desired inequality. (The inequality to be proved can be strict because as $x$ ranges over $A \cup B$, $b$ doesn’t necessarily range over all of $B$.)
   
   (b) Does the inequality $R_n^A(A \cup B) \leq R_n^A(A) + R_n^B(B)$ hold if it is assumed $0 \in A$ (but not necessarily $0 \in B$)? Justify your answer with either a proof or counter example.
   
   **Solution:** No, for example, if $n = 1$, $A = \{0, 1\}$ and $B = \{-1\}$ then $R_n^A(A \cup B) = 1$ and $R_n^0(A) = 1/2$ and $R_n^1(B) = 0$.
   
   (c) Suppose $n = 2$, $A = \{(x, x) : -1 \leq x \leq 1\}$, $\varphi_1(x_1) = x_1$, $\varphi_2(x_2) = |x_2|$. Find $R_2^A(A)$, $R_2^B(\varphi \circ A)$, $R_2(A)$, $R_2(\varphi \circ A)$. Do the contraction principle bounds hold with equality?
   
   **Solution:** Both $\varphi_1$ and $\varphi_2$ are 1-Lipschitz continuous and $\varphi(0) = 0$. The definitions yield
   
   - $\varphi \circ A = \{(x, |x|) : -1 \leq x \leq 1\}$,
   - $A \cup (-A) = A$,
   - $\varphi \circ A \cup (-\varphi \circ A) = \{(x, x) : -1 \leq x \leq 1\} \cup \{(x, -x) : -1 \leq x \leq 1\}$.
   - $R_2^A(A) = R_2^B(\varphi \circ A) = 1/2$,
   - $R_2(A) = 1/2$ and $R_2(\varphi \circ A) = 1$.
   
   Both contraction principle bounds hold with equality for this example.

2. **[VC dimension of combined classifiers using hard thresholding]**
   Let $\mathcal{G}$ denote the set of classifiers $g : \mathbb{R} \to \{-1, 1\}$ of the form $g_c(x) = \text{sgn}(x - c) = 1_{\{x \geq c\}} - 1_{\{x \leq c\}}$.
   
   (a) What is the VC dimension, $V(\mathcal{G})$? Also, what is the resulting upper bound on the Rademacher average, $R_n(\mathcal{G}(x^n))$, valid for all $x^n \in \mathbb{R}^n$ provided by Theorem 7.1 in the notes (which is based on the finite class lemma for Rademacher averages and the Sauer-Shelah lemma)?
   
   **Solution:** $V(\mathcal{G}) = 1$ so $R_n(\mathcal{G}(x^n)) \leq 2 \sqrt{\frac{\log(n+1)}{n}}$.
   
   (b) Let $\mathcal{G}_1$ be the set of classifiers of the form $g(x) = \text{sgn} \left( \sum_{i=1}^{N} c_i g_i(x) \right)$, for $N \geq 1$, $g_i \in \mathcal{G}$ for $i \in [N]$, and $c^N \in \mathbb{R}^N$ with $|c_1| + \ldots + |c_N| \leq 1$, or equivalently, $\mathcal{G}_1 = \text{sgn}(\text{absconv}(\mathcal{G}))$. Identify the VC dimension of $\mathcal{G}_1$ and the maximum Rademacher average for $n$ sample points, $R_n(\mathcal{G}_1(x^n))$.
   
   **Solution:** Given any $N \geq 1$ and $a < b$, let $g_{[a, b)}(x) = (\text{sgn}(x-a) - \text{sgn}(x-b))/(2N) = 1_{\{x \in [a, b)\}}/N$. Then functions of the form $\sum_{i=1}^{N} c_i g_{[a_i, b_i)}$ for $a_1 < b_1 < a_2 < b_2 < \ldots < b_N$ take values $\pm 1$ in each of the $N$ disjoint intervals $[a_i, b_i]$ and these functions are in $\mathcal{G}_1$. Thus, $\mathcal{G}_1$ shatters the set of $N$ midpoints of these intervals. Hence $V(\mathcal{G}_1) = \infty$. Also, $R_n(\mathcal{G}_1(x^n)) = 1$ for all $n \geq 1$ for any $x^n = \{x_1, \ldots, x_n\} \subseteq \mathbb{R}$ with cardinality $n$, because $\mathcal{G}_1(x^n) = \{-1, 1\}^n$. 
   
   Due beginning of class, Tuesday, March 5
3. [Bayesian estimators]

A Bayesian estimation problem of statistical decision theory can be represented by \((X, Y, U, P, \ell)\). As in the setup for model-free learning, \(X\) is a space of features, \(Y\) is a space of labels, \(U\) is a space of decision values, \(P\) is a known joint probability distribution on \(Z = X \times Y\) and \(\ell : Y \times U \to \mathbb{R}_+\) is a loss function. In contrast to model-free learning, only one probability distribution \(P\) is given and known, and a family of hypotheses \(\mathcal{F}\) is not needed. The Bayesian estimation problem is to find a function \(g : X \to U\) to minimize the expected loss (computed under \(P\)): \(\mathbb{E} [\ell(Y, g(X))].\) (If the function \(g\) were constrained to come from some function class \(\mathcal{F}\) it would be called an inductive bias constraint.)

(a) Show that the expected loss is minimized by \(g^*(x) = \arg \min_{u \in U} \mathbb{E} [\ell(Y, u)|X = x].\) (Hint: Use the tower property of conditional expectations, conditioning on \(X\) to compute \(\mathbb{E} [\ell(Y, g(X))].\))

**Solution:** The expected loss is an average of the conditional expected loss \(\mathbb{E} [\ell(Y, g(x))|X = x]\), which is minimized pointwise (i.e. for each \(x\)) by \(g^*(x)\). Therefore it minimizes the average as well. In more detail, for any \(g : X \to U\):

\[
\mathbb{E} [\ell(Y, g(X))] = \mathbb{E} [\mathbb{E} [\ell(Y, g(X))|X]] = \int_X \mathbb{E} [\ell(Y, g(X))|X = x] \mathbb{P} \{X \in dx\}
\]

(b) Give a simplified description of \(g^*\) in case \(Y = U = \mathbb{R}^d\) for some \(d \geq 1\) and \(\ell(s, y) = \|s - y\|^2\).

**Solution:** For any \(\mathbb{R}^d\) valued random vector \(V\) with \(\mathbb{E} [\|V\|^2] < \infty\) it can be shown that \(\arg \min_{v \in \mathbb{R}^d} \mathbb{E} [\|V - v\|^2] = \mu\), where \(\mu = \mathbb{E} [V]\). One proof is to expand \(\mathbb{E} [\|V - v\|^2]\) and set the gradient with respect to \(v\) to zero. Another proof is to use orthogonality:

\[
\mathbb{E} [\|V - v\|^2] = \mathbb{E} [\|V - \mu + \mu - v\|^2] = \mathbb{E} [\|V - \mu\|^2] + 2\mathbb{E} [(V - \mu, \mu - v)] + \|v - \mu\|^2
\]

\[
= \mathbb{E} [\|V - \mu\|^2] + 2\mathbb{E} [V - \mu, \mu - v] + \|v - \mu\|^2
\]

\[
= \mathbb{E} [\|V - \mu\|^2] + \|v - \mu\|^2 \geq \mathbb{E} [\|V - \mu\|^2].
\]

Applying this fact to the conditional distribution of \(Y\) given \(X = x\) for any given \(x\) (i.e. using part (a)) yields \(g^*(x) = E[Y|X = x]\). That is, \(g^*(X) = \mathbb{E} [Y|X]\) is the minimum mean squared error estimator.

(c) Give a simplified description of \(g^*\) in case \(Y\) has finite cardinality, \(Y = U\), and \(\ell(y, u) = 1\_{\{y \neq u\}}\).

**Solution:** For any \(Y\) valued random variable \(V\) with probability mass function \((p_V(i) : i \in Y)\),

\[
\arg \min_{i \in Y} \mathbb{E} [1_{\{V \neq i\}}] = \mathbb{arg} \min_{i \in Y} \mathbb{P} \{V \neq i\} = \mathbb{arg} \min_{i \in Y} (1 - p_V(i)) = \mathbb{arg} \max_{i \in Y} p_V(i).
\]

Applying this fact to the conditional probability distribution of \(Y\) given \(X = x\) yields \(g^*(x) = \mathbb{arg} \max_{i \in Y} \mathbb{P} \{Y = i|X = x\}\). This is the well known maximum posteriori probability (MAP) estimator.

4. [RKHS and inverse quadratic form]

Suppose that \(X = \{1, \ldots, n\} = [n]\). Functions on \(X\) are equivalent to column vectors, and the space of all functions on \(X\) is just \(\mathbb{R}^n\). Let \(K = (K_{i,j})_{i,j \in [n]}\) be a symmetric, positive definite matrix (so \(K\) is full rank). Then \(K\) is a Mercer kernel and the set of vectors in \(\mathcal{H}_K\), the column span of \(K\), is \(\mathbb{R}^n\). Let \(e_1, \ldots, e_n\) denote the standard orthonormal basis for \(\mathbb{R}^n\), so that \(e_i\) is the column vector with \(i^{th}\) coordinate equal to one and zeros elsewhere. Let \(Q_{i,j} = (e_i, e_j)_{K}\) for \(i, j \in [n]\).
5. [Half-space classifiers and support vector machines (SVM)]

Consider the concept learning problem \((X, Y) = \mathbb{R}^d, \{\pm 1\}, \mathcal{P}, \mathcal{G}\) with 0-1 loss, where \(\mathcal{P}\) is a set of probability distributions \(P\) on \(Z = X \times \{\pm 1\}\), and \(\mathcal{G}\) consists of all half-space classifiers of the form \(g_{w, b}(x) = \text{sgn}(\langle w, x \rangle + b)\), where \(w \in \mathbb{R}^d\) and \(b \in \mathbb{R}\). The generalization loss is defined by \(L_P(w, b) = P\{Y \neq \text{sgn}(\langle w, X \rangle + b)\}\).

(a) Explain why this problem is PAC learnable. That is, describe a PAC learning algorithm and give a performance bound demonstrating PAC learnability. (Hint: The set of classifiers considered is a Dudley class. The bound you give must depend on \(d\). Below we find a bound that does not depend on \(d\) in the realizable case, under a restriction on the width of the margin.)

Solution: (This is done in Section 8.1.1 of the notes.) The set of classifiers is a Dudley class based on a linear space of dimension \(d + 1\), so \(V(\mathcal{G}) = d + 1 < \infty\). Thus, the problem is PAC learnable by the fundamental theorem for PAC learnability of classification problems. Moreover, there is a universal constant \(C\) such that, given any \(\delta \in (0, 1)\), if \((w_n, b_n)\) is any ERM classifier (minimizing the number of mislabeled samples; it is not unique) then with probability at least \(1 - \delta\),

\[
L_P(w_n, b_n) \leq L_P(\mathcal{G}) + C \sqrt{\frac{d + 1}{n}} + \frac{2 \log(1/\delta)}{n}.
\]

(b) Given \(x \in \mathbb{R}^d\) and a classifier \((w, b)\) with \(w \neq 0\), let \(\pi(x)\) denote the projection of \(x\) onto the hyperplane defined by \(\langle w, x \rangle + b = 0\). Express \(\pi(x)\) and the distance, \(\|x - \pi(x)\|\), between \(x\) and the hyperplane in terms of \(x, w, b\), and \(b\). (Hint: Since \(w\) is normal to the hyperplane, \(\pi(x)\) is the point in the hyperplane of the form \(\pi(x) = x - cw\) for some constant \(c\)).

Solution: Following the hint, we find \(c\) so that \(\langle x - cw, w \rangle + b = 0\), or \(\langle x, w \rangle - c\|w\|^2 + b = 0\), yielding \(\pi(x) = x - \left\{\frac{\langle w, x \rangle + b}{\|w\|^2}\right\}w\) or, equivalently, \(\pi(x) = x - \left\{\frac{\langle w, x \rangle + b}{\|w\|^2}\right\}\frac{w}{\|w\|}\). Since \(\|w\| = 1\), we see the distance is equal to \(\|x - \pi(x)\| = \frac{\|\langle w, x \rangle + b\|}{\|w\|}\).

(c) Given a data set \(Z^n = ((X_1, Y_1), \ldots, (X_n, Y_n))\) and a classifier \((w, b)\) with \(w \neq 0\), let the margin, \(M_i\), of the \(i^{th}\) sample point be defined by \(M_i \triangleq Y_i((\langle w, X_i \rangle) + b)/\|w\|\). Thus, \(M_i\) is the signed distance of \(X_i\) from the hyperplane defined by \(\langle w, x \rangle + b = 0\), with the sign being positive if \(Y_i = \text{sgn}(\langle w, x_i \rangle + b)\) and negative otherwise. Define the margin for the whole data set by \(M \triangleq \min_{i \in [n]} M_i\). Suppose that \(M > 0\) for some choice of \((w, b)\). A key idea of SVMs is to find \((w, b)\) to maximize \(M\), with the hope that it will lead to a better classifier for fresh samples. Show that:

\[
\max_{(w, b)} M = \max\left\{\frac{1}{\|w\|} : (w, b) \text{ subject to } Y_i((\langle w, x_i \rangle) + b) \geq 1 \text{ for } i \in [n]\right\}
\]
(Hint: $M$ for a given $(w, b)$ is not changed if $(w, b)$ is multiplied through by a positive scaler.)

Remark: The right-hand side of (2) represents an optimization problem that is equivalent to the quadratic optimization problem (4) below.

**Solution:** By the hint, in the maximization problem on the left-hand side of (2), we can impose the constraint that $\min_{i \in [n]} Y_i((w, X_i) + b) = 1$. Under that constraint, $M = \frac{1}{||w||}$. Therefore,

$$\max_{(w, b)} M = \max \left\{ M : (w, b) \text{ subject to } \min_{i \in [n]} Y_i((w, X_i) + b) = 1 \right\}$$

$$= \max \left\{ \frac{1}{||w||} : (w, b) \text{ subject to } \min_{i \in [n]} Y_i((w, X_i) + b) = 1 \right\}$$

(3)

For the final step, note that the right-hand sides of (2) and (3) are equal, because if $(w, b)$ satisfies all the inequalities in (2) with strict inequality, then they would continue to hold if $(w, b)$ were multiplied by a scaler slightly less than one, thereby reducing $||w||$. (Note: A related approach to this problem is to consider the dual of the quadratic optimization problem (4).)

(d) (Bound not depending on $d$, realizable case with lower bound on relative margin) Suppose $C_K > 0$ and $\lambda > 0$. Let $P$ denote the set of all probability distributions $P$ on $Z = X \times \{ \pm 1 \}$ such that: $P(\sqrt{1 + ||X||^2} \leq C_K) = 1$, and there exists a classifier $(w, b)$ (depending on $P$) such that $||w||^2 + b^2 \leq \lambda^2$ and $P(Y((w, X) + b) \geq 1) = 1$. These assumptions ensure that iid samples generated by $P$ satisfy the following with probability one: $||X|| \leq C_K$ for each $i$, and there exits $(w, b)$ for the data points with margin $M$ at least $1/\lambda$. Thus, the ratio of the margin to $max_i ||X||$ is greater than or equal to $\frac{1}{\sqrt{\lambda}}$. Of course, just because the data samples can be separated by a particular hyperplane doesn’t necessarily mean that the hyperplane will classify fresh sample points well. Show that if $(\hat{w}_n, b_n)$ is the particular ERM classifier given by

$$(\hat{w}_n, b_n) = \arg \min \left\{ \frac{1}{n} : (w, b) \text{ subject to } Y_i((w, X_i) + b) \geq 1 \right\},$$

then with probability at least $1 - \delta$,

$$L_P((\hat{w}_n, b_n)) \leq \frac{4\lambda C_K}{\sqrt{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}.$$  

(5)

The bound (5) does not depend on the dimension, $d$, of the feature space. (Hint: Bring in a Mercer kernel $K$ and apply Corollary 8.2.)

**Solution:** Let $\varphi$ be the ramp penalty function, $\varphi(x) = \min\{1, 1 + x\}$. The assumptions imply that, for the learning algorithm (4), the surrogate empirical loss is zero: $A_{\varphi,n}(\hat{w}_n, b_n) = 0$. Let $K$ denote the Mercer kernel on $X \times X$ defined by $K(x, x') = 1 + \langle x, x' \rangle$, and let $\mathcal{H}_K$ denote the associated RKHS. The set of classifiers we consider is $\mathcal{F}_\lambda$, the zero-centered ball of radius $\lambda$ in $\mathcal{H}_K$, so that $E[R_n(\mathcal{F}_\lambda(X^n))] \leq \lambda \sqrt{\frac{\mathbb{E}[K(X, X)]}{n}} \lambda \sqrt{\mathbb{E}[1 + \|X\|^2]} \leq \frac{\lambda C_K}{\sqrt{n}}$. So (5) follows from Corollary 8.2 with $\varphi$ equal to the ramp function. (ACTUALLY, IT DOESN’T FOLLOW BECAUSE WE DON’T KNOW THAT THE CLASSIFIER CHOSEN SATISFIES $\|w\|^2 + b^2 \leq \lambda$. THANKS TO FOREST FOR POINTING OUT ERROR. THIS PART WORKS IF $\|w\|^2$ IN (4) IS REPLACED BY $\|W\|^2 + b^2$. THIS IS EQUIVALENT TO MAXIMIZING THE MARGIN AFTER ADDING AN ADDITIONAL COORDINATE TO EACH DATA POINT THAT IS EQUAL TO ONE FOR EACH DATA POINT AND THINKING OF THE MARGIN IN $\mathbb{R}^{d+1}$. IF SUCH CHANGE TO THE DATA WERE DONE AT THE BEGINNING OF THE PROBLEM WE COULD LEAVE OUT $b$ ALTOGETHER.)

6. [A view of AdaBoost in SK learn]

The python programming problem for this problem set is explained within the .ipynb file. You can see a static version at http://nbviewer.jupyter.org/urls/courses.engr.illinois.edu/ee543/sp2019/ee543_PythonProblem3.ipynb?flush_cache=true and download the ipynb file from the static version or directly from https://courses.engr.illinois.edu/ee543/sp2019/ee543_PythonProblem3.ipynb.
Solution: (a) The choice of weight vector puts more weight on data samples that have a large absolute value of one of the particular features (i.e. one of the coordinates of the feature vector). This steers the decision tree classifier to focus more on those data points. Values of that feature apparently give more informative for those data samples than the other features. That is why we see that the top level split of the decision tree is in most cases based on the values of that feature. (b) Increasing the depth of the dt classifier helps the algorithm train more quickly on the training data, but the performance on the test data is no better. (c) Smaller values of the learning rate slow the rate of training on the training data, but it doesn’t seem to help with the test data.