Please visit the course website soon: http://courses.engr.illinois.edu/ece543/

1. Preliminaries and some basic tools

Assigned reading: Chapters 1-3 of Prof. Raginsky’s notes, available at https://courses.engr.illinois.edu/ece543/SLT_Jan17_2017.pdf. (Introduction, Concentration Inequalities, and Formulation of the Learning Problem) http://maxim.ece.illinois.edu/teaching/fall15b/schedule.html. Later sections of the notes might be updated later in the semester, so it might not be a good idea to print out the entire set.

Additional recommended reading: Shalev-Shwartz and Ben-David, Understanding Machine Learning, from Theory to Algorithms, Chapters 1-3 and Appendix B.

Problems to be handed in:

1. [PAC learnability of right subintervals]
   Consider the concept learning problem for the triple \((X, \mathcal{P}, \mathcal{C})\), where \(X = [0, 1]\), \(\mathcal{C} = \{\tau, 1\} : 0 \leq \tau \leq 1\), and \(\mathcal{P}\) is the set of all probability distributions on \(X\). Describe an ERM classifier \(\hat{C}_n\) and show that it PAC learns \(\mathcal{C}\) in the realizable case to accuracy \(\epsilon\) with probability at least \(1 - \delta\), (i.e.\)
   \[
   P\{P(C^* \triangle \hat{C}) \leq \epsilon\} \geq 1 - \delta, \tag{1}
   \]
   if the sample size is at least \(n(\epsilon, \delta) = \left\lceil \frac{\log(1/\delta)}{\epsilon^2} \right\rceil\). The training set is given by \(Z^n = (Z_1, \ldots, Z_n)\), where \(Z_i = (X_i, Y_i) = (X_i, 1_{\{X_i \in C^*\}})\) and \(C^* = [\tau^*, 1]\) is the target concept.

2. [The minimum of a uniform approximation is an approximate minimizer]
   Suppose we’d like to find a minimizer of a function \(G\) defined on some domain \(\mathcal{U}\), but the function \(G\) is not known. Suppose that \(\hat{G}\) is an \(\epsilon\) uniform approximation of \(G\) for some \(\epsilon > 0\), meaning that \(|G(u) - \hat{G}(u)| \leq \epsilon\) for all \(u \in \mathcal{U}\). Suppose that \(u^*\) is a minimizer of \(G\), meaning that \(u^* \in \mathcal{U}\) and \(G(u^*) \leq \hat{G}(u^*)\) for all \(u \in \mathcal{U}\). Prove that \(G(u^*) \leq \inf_{u \in \mathcal{U}} G(u) + 2\epsilon\).

3. [Subgaussian random variables]
   A random variable \(X\) is said to be subgaussian with scale parameter \(\nu\), if \(X\) has a finite mean and \(\mathbb{E}[e^{sX}] \leq e^{\frac{s^2}{2} \nu^2}\) for all \(s \in \mathbb{R}\). Sometimes \(\nu^2\) is called the proxy variance because a Gaussian random variable with variance \(\sigma^2\) is subgaussian for \(\nu = \sigma\).
   (a) Suppose \(U\) is a random variable such that for some parameters \(a, b, \mathbb{P}\{U \in [a, b]\} = 1\). What is the smallest value of \(\nu\), depending only on \(a\) and \(b\), for which it follows that \(U\) is subgaussian with scale parameter \(\nu^2\)? (Hint: Consider the Hoeffding inequality.)
   (b) Suppose \(X\) has variance \(\sigma^2\) and proxy variance \(\nu^2\). Is it true that \(\sigma^2 \leq \nu^2\)? (Hint: What is the power series expansion of \(\cosh(sx)\) with respect to \(x\) about zero?)
   (c) Suppose \(S_n = X_1 + \cdots + X_n\), where \(X_1, \ldots, X_n\) are independent random variables, such that \(X_i\) is subgaussian with scale parameter \(\nu_i\). Show that \(S_n\) is subgaussian with proxy variance given by \(\nu^2 = \nu_1^2 + \cdots + \nu_n^2\). Does this fact continue to hold if the \(X\)’s are not independent? Justify your answer with either a proof or counter example.
   (d) Using the methodolgy of the Chernoff inequality, show that if \(X\) is subgaussian with scale parameter \(\nu\), then \(\mathbb{P}\{X - \mathbb{E}[X] \geq nt\} \leq e^{-t^2 \nu^2}\) for all \(t \geq 0\).
   (e) Suppose \(X_1, \ldots, X_n\) are mean zero, and each is subgaussian with scale parameter \(\nu\). Show that \(\mathbb{E}[\max_i X_i] \leq \nu\sqrt{2\log n}\). Does this bound require the \(X\)’s to be independent? (Hint: Use Jensen’s inequality for \(e^{x}\), and the inequality \(e^{sx}X_i \leq \sum_s e^{sx}X_i\) for \(s \geq 0\).)
   (f) Under the same assumptions as the previous part, show that \(\mathbb{P}\{\max_i X_i \geq \nu(\sqrt{2\log n} + t)\} \leq e^{-t^2\nu^2} e^{-\nu t^2 / 2}\). (Hint: Use the union bound for probabilities and part (d).)

4. [Bounded differences vs. Lipschitz continuity]
   Fix \(p > 0\) and define \(f : [0, 1]^n \to \mathbb{R}_+\) by \(f(x) = \|x\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}\). If \(p \geq 1\) then \(\|x\|_p\) is a norm called the \(\ell^p\) norm. Suppose \(n \geq 2\).
(a) Find the smallest constant $c > 0$ so that $f$ has the bounded differences property (as appearing in McDiarmid’s inequality). Your answer may depend on $n$ and $p$. Check your answers for $p = 1$ and $p = 2$.

(b) Find the smallest constant $L > 0$ such that $f$ is $L$-Lipschitz continuous (meaning $|f(x) - f(y)| \leq L\|x - y\|$). Your answers may depend on $n$ and $p$. Hint: For continuously differentiable functions on convex sets, the smallest $L$ is the max of $\|\nabla f\|$ over the domain of $f$. If no finite $L$ works, then $f$ is not Lipschitz continuous.

5. **Convexity and $f$ divergences**

(We use notation $\mathbb{R}_+ = [0, +\infty)$ and $\mathbb{R}_+^2 = (0, +\infty)^2$.) Let $f$ be a finite convex function on $\mathbb{R}_+$ such that $f(1) = 0$ and $\lambda f(a) + \lambda f(b) > f(1)$ whenever $0 < a < 1 < b$, $\lambda \in (0, 1)$, and $\lambda a + \lambda b = 1$, where $\lambda = 1 - \lambda$ (i.e. $f$ is strictly convex at one).

(a) Show that $bf\left(\frac{x}{y}\right)$ is a convex function of $(a,b)$ over $\mathbb{R}_+^2$ (i.e. it is jointly convex in $a$ and $b$).

(b) Show that the function $(a,b) \mapsto bf\left(\frac{x}{y}\right)$ can be uniquely extended to a function $\varphi$ on $\mathbb{R}_+^2 \equiv [0, \infty)^2$ with values in $\mathbb{R} \cup \{+\infty\}$ such that: $\varphi(0,0) = 0$ and $\varphi$ is continuous on $\mathbb{R}_+^2 \setminus \{(0,0)\}$. Also, show that the extension $\varphi$ is convex over all of $\mathbb{R}_+^2$. (Hint: So $\varphi(a,b) = bf\left(\frac{x}{y}\right)$ for $(a,b) \in \mathbb{R}_+^2$ and $\varphi(0,0) = 0$. It remains to define $\varphi$ at the nonzero boundary points, i.e. at points of the form $(a_n,0)$ with $a_n > 0$ or $(0,b_n)$ with $b_n > 0$, in such a way that the resulting function is continuous at those boundary points. The following fact is relevant: If $f$ is a convex function over $(0, \infty)$ then the limits $\lim_{r \to +\infty} \frac{1}{r}f(r)$ and $\lim_{r \to 0} f(r)$ exist, with values in $\mathbb{R} \cup \{+\infty\}$.)

(c) The $f$-divergence between probability vectors $p$ and $q$ (with the same dimension) is defined by

$$D_f(p||q) = \sum_i \varphi(p_i,q_i).$$

Important examples for $f(u)$ are

<table>
<thead>
<tr>
<th>$f(u)$</th>
<th>name of $f$ divergence</th>
<th>expression for $f$ divergence</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u \log u$</td>
<td>Kullback-Liebler (KL) divergence</td>
<td>$D(p</td>
</tr>
<tr>
<td>$-\log u$</td>
<td>KL divergence with $p$ and $q$ reversed</td>
<td></td>
</tr>
<tr>
<td>$\frac{1}{2}</td>
<td>u-1</td>
<td>$</td>
</tr>
<tr>
<td>$(\sqrt{u} - 1)^2$</td>
<td>squared Hellinger distance</td>
<td>$H(p,q)^2 = \sum_i (\sqrt{p_i} - \sqrt{q_i})^2$</td>
</tr>
<tr>
<td>$(u - 1)^2$</td>
<td>$\chi^2$ divergence</td>
<td>$\chi^2(p</td>
</tr>
</tbody>
</table>

Show that $D_f(p||q)$ is jointly convex in $p$ and $q$, and that $D_f(p||q) \geq 0$, with equality if and only if $p = q$. (Hint: See Jensen’s inequality and the proof of Jensen’s inequality.)

6. **Getting started on Python**

Suppose the $k$ nearest neighbor (KNN) classifier is used for classification of the classic Iris flower data set, with uniform weighting of $k$ nearest neighbors. Suppose the 150 samples are randomly, uniformly divided into 75 training samples, used to train the classifier, and 75 samples to test the classifier. The scores on the training set and on the test set are defined to be the fraction of correct classifications for those sets. The scores are random because of the random splitting of the samples into two groups. Here is what you need to compute: (a) Determine the number of neighbors $k$ that maximizes the expected score of the trained classifier on the training sample. (b) Determine the number of neighbors $k$ that maximizes the expected score of the trained classifier on the test data. To get started, see [http://nbviewer.jupyter.org/urls/courses.engr.illinois.edu/ece543/ece543_PythonProblem1.ipynb?flush_cache=true](http://nbviewer.jupyter.org/urls/courses.engr.illinois.edu/ece543/ece543_PythonProblem1.ipynb?flush_cache=true).