This homework is due at 23:59, April 14, 2019.

Problem 1

In this problem, we will derive a generalization error bound for the *support vector machine*.

Consider the binary classification problem. Let $(x, y), (x_1, y_1), \dots, (x_n, y_n)$ be independent and identically distributed (i.i.d.) random variables in $\mathscr{X} \times \{\pm 1\}$ for some set $\mathscr{X}$. Let $\mathscr{H}$ be a class of hypotheses $h : \mathscr{X} \to [-B, B]$ for some $B > 0$. Suppose for any given $x \in \mathscr{X}$, we predict the corresponding y by the sign of $h(x)$. Consider the $0-1$ loss

$$\lambda(u) := \begin{cases} 1 & \text{, if } u \le 0, \\ 0 & \text{, otherwise,} \end{cases}$$

and the corresponding risk and empirical risk functions

$$R(h) := \mathsf{E}\, \lambda(y h(x)), \quad R_n(h) := \frac{1}{n} \sum_{i=1}^{n} \lambda(y_i h(x_i)), \quad \forall h \in \mathscr{H}.$$

Define the hinge loss

$$\varphi(u) := \max\{0, 1 - u\}, \quad \forall u \in \mathbb{R}.$$

The support vector machine outputs a hypothesis $\hat{h}_n$ that minimizes the average hinge loss

$$\Phi_n(h) := \frac{1}{n} \sum_{i=1}^{n} \varphi(y_i h(x_i)), \forall h \in \mathscr{H},$$

on the hypothesis class $\mathscr{H}$.

1. (10 points) We say a function $\psi : \mathbb{R} \to \mathbb{R}$ is a *contraction*, if and only if

$$|\psi(y) - \psi(x)| \le |y - x|, \quad \forall x, y \in \mathbb{R}.$$

Theorem 1 (Contraction principle [2])**.** *Let $\mathscr{A} \subset \mathbb{R}^n$. Let $\psi : \mathbb{R} \to \mathbb{R}$ be a contraction. Define*

$$\psi \circ \mathscr{A} := \left\{ (\psi(a_1), \dots, \psi(a_n)) \,\middle|\, (a_1, \dots, a_n) \in \mathscr{A} \right\}.$$

Then, it holds that

$$\mathsf{E}\left[\sup_{(b_1,\dots,b_n) \in \psi \circ \mathscr{A}} \sum_{i=1}^{n} \sigma_i b_i \right] \le \mathsf{E}\left[\sup_{(a_1,\dots,a_n) \in \mathscr{A}} \sum_{i=1}^{n} \sigma_i a_i \right],$$

where $\sigma_1, \dots, \sigma_n$ are i.i.d. Rademacher random variables.

Let $\mathscr{F}$ be the set $\{ f_h : (x, y) \to \varphi(y h(x)) \mid h \in \mathscr{H} \}$. **Use Theorem 1 to show that**

$$\hat{C}_n(\mathscr{F}) \le \hat{C}_n(\mathscr{H}), \tag{1}$$

where $\hat{C}_n$ denotes the empirical Rademacher complexity, i.e.,

$$\hat{C}_n(\mathscr{F}) := \mathsf{E}_{\sigma_1,\dots,\sigma_n} \sup_{h \in \mathscr{H}} \left[\frac{1}{n} \sum_{i=1}^{n} \sigma_i f_h(x_i, y_i) \right], \quad \hat{C}_n(\mathscr{H}) := \mathsf{E}_{\sigma_1,\dots,\sigma_n} \sup_{h \in \mathscr{H}} \left[\frac{1}{n} \sum_{i=1}^{n} \sigma_i h(x_i) \right],$$

with $\sigma_1, \dots, \sigma_n$ being i.i.d. Rademacher random variables.

Solution. Obviously, φ is a contraction. Let

$$\mathscr{A} := \{ (y_1 h(x_1), \dots, y_n h(x_n)) \in \mathbb{R}^n \mid h \in \mathscr{H} \}.$$

Problem 1 continued on the next page. . .

Then, by Theorem 1, we write

$$\begin{aligned}
\hat{C}_n(\mathscr{F}) &= \mathsf{E}\left[\sup_{(b_1,\dots,b_n)\in\varphi\circ\mathscr{A}} \sum_{i=1}^n \sigma_i b_i\right] \\
&\leq \mathsf{E}\left[\sup_{(a_1,\dots,a_n)\in\mathscr{A}} \sum_{i=1}^n \sigma_i a_i\right] \\
&= \mathsf{E}_{\sigma_1,\dots,\sigma_n} \sup_{h\in\mathscr{H}} \left[\frac{1}{n}\sum_{i=1}^n \sigma_i y_i h(x_i)\right] \\
&= \mathsf{E}_{\sigma_1,\dots,\sigma_n} \sup_{h\in\mathscr{H}} \left[\frac{1}{n}\sum_{i=1}^n \sigma_i h(x_i)\right] \\
&= \hat{C}_n(\mathscr{H}).
\end{aligned}$$

Notice the second last equality holds because $y_i \in \{\pm 1\}$.

2. (10 points) **Use** (1) **to show that for any** $\delta \in]0,1[$**,**

$$\mathsf{P}\left(R(\hat{h}_n) \leq \Phi_n(\hat{h}_n) + 2C_n(\mathscr{H}) + (B+1)\sqrt{\frac{\log(1/\delta)}{2n}} \right) \geq 1-\delta,$$

where C_n **denotes the Rademacher complexity, i.e.,**

$$C_n(\mathscr{H}) := \mathsf{E}_{(x_1,y_1),\dots,(x_n,y_n)} \hat{C}_n(\mathscr{H}).$$

Solution. Define the expected hinge loss

$$\Phi(h) := \mathsf{E}\,\varphi(yh(x)), \quad \forall h \in \mathscr{H}.$$

Notice that

$$0 \leq \varphi(yh(x)) \leq B+1, \quad \forall (x,y) \in \mathscr{X}\times\{\pm 1\}, h \in \mathscr{H}.$$

Then, applying Theorem 1 of Lecture 2 with the normalized hinge loss $\varphi/(B+1)$, we write for any $\delta \in]0,1[$,

$$\mathsf{P}\left(\Phi(\hat{h}_n) \leq \Phi_n(\hat{h}_n) + 2C_n(\mathscr{F}) + 2\sqrt{\frac{\log(1/\delta)}{2n}} \right) \geq 1-\delta.$$

As

$$\lambda(yh(x)) \leq \varphi(yh(x)), \quad \forall (x,y) \in \mathscr{X}\times\{\pm 1\}, h \in \mathscr{H},$$

we have $\Phi(\hat{h}_n) \geq R(\hat{h}_n)$. It remains to apply (1).

Problem 2

In this problem, we will derive a PAC Bayesian-type generalization error bound for a countable hypothesis class.

Let $z, z_1, \dots, z_n$ be i.i.d. random variables taking values in a set $\mathcal{Z}$. Let $\mathscr{H}$ be a countable set of hypotheses $h: \mathcal{Z} \to \mathbb{R}$. Let $\lambda: \mathscr{H}\times\mathcal{Z} \to [0,1]$ be a bounded loss function. Define

$$R(h) := \mathsf{E}_z\,\lambda(h,z), \quad R_n(h) := \frac{1}{n}\sum_{i=1}^n \lambda(h,z_i), \quad \forall h \in \mathscr{H}.$$

Let π and $\hat{\pi}$ be two probability distributions on $\mathscr{H}$. Suppose π is independent of $z, z_1, \dots, z_n$

Problem 2 continued on the next page...

1. (10 points) **Show that for any** $\delta \in]0,1[$**,**

$$\mathsf{P}\left(\mathsf{E}_{\hat{h}\sim\pi} R(\hat{h}) \leq \mathsf{E}_{\hat{h}\sim\pi} R_n(\hat{h}) + \sqrt{\frac{H(\pi)+\log(1/\delta)}{2n}} \right) \geq 1-\delta,$$

where H **denotes the entropy function, i.e.,**

$$H(\pi) := -\sum_{h\in\mathcal{H}} \pi(h)\log\pi(h).$$

HINT: Recall the following result in Lecture 4:

$$\mathsf{P}\left(\forall h \in \mathcal{H} : R(h) \leq R_n(h) + \sqrt{\frac{\log(1/\pi(h))+\log(1/\delta)}{2n}} \right) \geq 1-\delta.$$

Solution. By the hint, we have

$$\mathsf{P}\left(\mathsf{E}_{\hat{h}\sim\pi} [R(h) - R_n(h)] \leq \mathsf{E}_{\hat{h}\sim\pi} \sqrt{\frac{\log(1/\pi(\hat{h}))+\log(1/\delta)}{2n}} \right) \geq 1-\delta.$$

By Jensen's inequality, we obtain

$$\mathsf{P}\left(\mathsf{E}_{\hat{h}\sim\pi} [R(h) - R_n(h)] \leq \sqrt{\frac{\mathsf{E}_{\hat{h}\sim\pi}\left[\log(1/\pi(\hat{h}))\right]+\log(1/\delta)}{2n}} \right) \geq 1-\delta.$$

It remains to notice that

$$\mathsf{E}_{\hat{h}\sim\pi}\left[\log\left(\frac{1}{\pi(\hat{h})}\right)\right] = H(\pi).$$

2. (10 points) **Show that for any** $\delta \in]0,1[$**,**

$$\mathsf{P}\left(\mathsf{E}_{\hat{h}\sim\hat{\pi}} R(\hat{h}) \leq \mathsf{E}_{\hat{h}\sim\hat{\pi}} R_n(\hat{h}) + \sqrt{\frac{D(\hat{\pi}\|\pi)+H(\hat{\pi})+\log(1/\delta)}{2n}} \right) \geq 1-\delta,$$

where D **denotes the relative entropy, i.e.,**

$$D(\hat{\pi}\|\pi) := \sum_{h\in\mathcal{H}} \hat{\pi}(h)\log\frac{\hat{\pi}(h)}{\pi(h)}.$$

Solution. Similarly as in the derivation above, we write

$$\mathsf{P}\left(\mathsf{E}_{\hat{h}\sim\hat{\pi}} [R(h) - R_n(h)] \leq \sqrt{\frac{\mathsf{E}_{\hat{h}\sim\hat{\pi}}\left[\log(1/\pi(\hat{h}))\right]+\log(1/\delta)}{2n}} \right) \geq 1-\delta.$$

It remains to notice that

$$\mathsf{E}_{\hat{h}\sim\hat{\pi}}\left[\log\left(\frac{1}{\pi(\hat{h})}\right)\right] = \mathsf{E}_{\hat{h}\sim\hat{\pi}}\left[\log\left(\frac{\hat{\pi}(\hat{h})}{\pi(\hat{h})}\right) - \log\left[\hat{\pi}(\hat{h})\right]\right] = D(\hat{\pi}\|\pi) + H(\hat{\pi}).$$

3. (10 points) **Explain why we require** π **to be independent of** $z, z_1, \ldots, z_n$**, while we do not require** $\hat{\pi}$ **to satisfy the same condition.**

Solution. If π is not independent of $z, z_1, \ldots, z_n$, the inequality in the hint may not hold.

Problem 3

In this problem, we will derive a general inequality that can yield a variety of PAC Bayesian bounds, and a special case of it.

Consider the same setting as in Problem 2, except that now the hypothesis class is general and not necessarily countable.

1. (20 points) Let $\varphi : \mathscr{H} \to \mathbb{R}$ possibly dependent on $z_1, \ldots, z_n$. The *change of measure inequality* says that for any $\eta \in]0, +\infty[$,
$$\eta \mathsf{E}_{\hat{h}\sim\hat{\pi}}\, \varphi(\hat{h}) \le D(\hat{\pi}\|\pi) + \log\left(\mathsf{E}_{\hat{h}\sim\pi} \mathrm{e}^{\eta\varphi(\hat{h})}\right),$$
where $D(\hat{\pi}\|\pi)$ denotes the relative entropy between $\hat{\pi}$ and π. **Use the inequality to show that for any $\eta \in]0, +\infty[$ and $\delta \in]0, 1[$,**
$$\mathsf{P}\left(\mathsf{E}_{\hat{h}\sim\hat{\pi}}\varphi(\hat{h}) \le \frac{1}{\eta} D(\hat{\pi}\|\pi) + \frac{1}{\eta}\log\frac{C_n(\eta)}{\delta}\right) \ge 1-\delta, \tag{2}$$
where the probability is with respect to the randomness of $z_1, \ldots, z_n$, and
$$C_n(\eta) := \mathsf{E}_{z_1,\ldots,z_n} \mathsf{E}_{\hat{h}\sim\pi} \mathrm{e}^{\eta\varphi(\hat{h})}.$$

   Remark. Rigorously speaking, the term $\mathsf{E}_{\hat{h}\sim\hat{\pi}}\, \varphi(\hat{h})$ in the change of measure inequality should be understood as
$$\mathsf{E}_{\hat{h}\sim\hat{\pi}}\, \varphi(\hat{h}) := \sup_{B\in\mathbb{R}} \mathsf{E}_{\hat{h}\sim\hat{\pi}} \min\left\{B, \varphi(\hat{h})\right\}.$$
See [1, Section 5.2] for the details. *You can ignore this mathematical subtlety for this homework.*

   Solution. By Markov's inequality, we write
$$\mathsf{P}\left(\mathsf{E}_{\hat{h}\sim\pi} \mathrm{e}^{\eta\varphi(\hat{h})} \le \frac{C_n(\eta)}{\delta}\right) \ge 1-\delta.$$
Equivalently, we have
$$\mathsf{P}\left(\log \mathsf{E}_{\hat{h}\sim\pi} \mathrm{e}^{\eta\varphi(\hat{h})} \le \log\frac{C_n(\eta)}{\delta}\right) \ge 1-\delta.$$
By the change of measure inequality, we obtain
$$\mathsf{P}\left(\eta \mathsf{E}_{\hat{h}\sim\hat{\pi}}\, \varphi(\hat{h}) - D(\hat{\pi}\|\pi) \le \log\frac{C_n(\eta)}{\delta}\right) \ge 1-\delta,$$
the desired inequality.

2. (10 points) For any $p, q \in]0, 1[$, define
$$\delta(p\|q) := p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}.$$
Let u, v be random variables taking values in $]0, 1[$. **Show that**
$$\mathsf{E}\,\delta(u\|v) \ge \delta(\mathsf{E}\,u\|\mathsf{E}\,v). \tag{3}$$

   HINT: Let $\psi : \mathscr{X} \to \mathbb{R}$ for some set $\mathscr{X}$. *Jensen's inequality* says if for all $x, y \in \mathscr{X}$,
$$\psi((1-\alpha)x + \alpha y) \le (1-\alpha)\psi(x) + \alpha\psi(y), \quad \forall \alpha \in [0, 1], \tag{4}$$

Problem 3 continued on the next page...

then for any random variable ξ taking values in $\mathscr{X}$,

$$\mathsf{E}\,\psi(\xi) \geq \psi(\mathsf{E}\,\xi).$$

Solution. We write

$$\begin{aligned}\delta((1-\alpha)p+\alpha\tilde{p}\|(1-\alpha)q+\alpha\tilde{q}) = {}& \left[(1-\alpha)p+\alpha\tilde{p}\right]\log\left[(1-\alpha)p+\alpha\tilde{p}\right] + \\ & \left\{1-\left[(1-\alpha)p+\alpha\tilde{p}\right]\right\}\log\left\{1-\left[(1-\alpha)p+\alpha\tilde{p}\right]\right\} - \\ & \left[(1-\alpha)p+\alpha\tilde{p}\right]\log\left[(1-\alpha)q+\alpha\tilde{q}\right] - \\ & \left\{1-\left[(1-\alpha)p+\alpha\tilde{p}\right]\right\}\log\left\{1-\left[(1-\alpha)q+\alpha\tilde{q}\right]\right\}.\end{aligned}$$

Then, it suffices to check if both $v \mapsto v\log v$ and $u \mapsto -\log u$ satisfy (4); if yes, then $v \mapsto (1-v)\log(1-v)$ and $u \mapsto -\log(1-u)$ also satisfy (4) by symmetry. Let $v, \tilde{v} \in]0,1[$; assume $v \geq \tilde{v}$ without loss of generality. Then, we write

$$\begin{aligned}&[(1-\alpha)v+\alpha\tilde{v}]\log[(1-\alpha)v+\alpha\tilde{v}] - (1-\alpha)v\log v - \alpha\tilde{v}\log\tilde{v} \\ &\quad = (1-\alpha)v\log\frac{(1-\alpha)v+\alpha\tilde{v}}{v} + \alpha\tilde{v}\log\frac{(1-\alpha)v+\alpha\tilde{v}}{\tilde{v}} \\ &\quad \leq (1-\alpha)v\log\frac{(1-\alpha)v+\alpha\tilde{v}}{v} + \alpha v\log\frac{(1-\alpha)v+\alpha\tilde{v}}{\tilde{v}} \\ &\quad = \log\frac{(1-\alpha)v+\alpha\tilde{v}}{v} + v\log\frac{\tilde{v}}{v} \\ &\quad \leq 0.\end{aligned}$$

Checking if $u \mapsto -\log u$ satisfies (4) is easy, so we skip the proof.

3. (20 points) Below is a non-trivial result in probability theory.

Theorem 2 ([3])**.** *Let $\xi, \xi_1, \ldots, \xi_n$ be i.i.d. random variables taking values in $[0,1]$. Define*

$$\mu := \mathsf{E}\,\xi, \quad \hat{\mu}_n := \frac{1}{n}\sum_{i=1}^{n}\xi_i.$$

Then, it holds that

$$\mathsf{E}\,\mathrm{e}^{n\delta(\hat{\mu}_n\|\mu)} \leq 2\sqrt{n}, \quad \forall n \geq 8.$$

Assume $R(h) \in]0,1[$ for all $h \in \mathscr{H}$. **Use** (2), (3)**, and Theorem 2 to show that for any** $\delta \in]0,1[$ **and** $n \geq 8$**,**

$$\mathsf{P}\left(\delta\left(\mathsf{E}_{\hat{h}\sim\hat{\pi}}\,R_n(\hat{h})\,\middle\|\,\mathsf{E}_{\hat{h}\sim\hat{\pi}}\,R(\hat{h})\right) \leq \frac{1}{n}\left[D(\hat{\pi}\|\pi) + \log\frac{2\sqrt{n}}{\delta}\right]\right) \geq 1-\delta.$$

Solution. Set

$$\varphi(h) = \delta(R_n(h)\|R(h)).$$

Then, by (2) with $\eta = n$, we have

$$\mathsf{P}\left(\mathsf{E}_{\hat{h}\sim\hat{\pi}}\delta(R_n(\hat{h})\|R(\hat{h})) \leq \frac{1}{n}D(\hat{\pi}\|\pi) + \frac{1}{n}\log\frac{\tilde{C}_n}{\delta}\right) \geq 1-\delta,$$

where

$$\tilde{C}_n := \mathsf{E}_{z_1,\ldots,z_n}\mathsf{E}_{\hat{h}\sim\pi}\mathrm{e}^{n\delta(R_n(\hat{h})\|R(\hat{h}))}.$$

By Theorem 2, we write

$$\tilde{C}_n \leq \mathsf{E}_{\hat{h}\sim\pi}(2\sqrt{n}) = 2\sqrt{n}.$$

Problem 3 continued on the next page. . .

Then, we obtain

$$\mathsf{P}\left(\mathsf{E}_{\hat{h}\sim\hat{\pi}}\delta(R_n(\hat{h})\,\|\,R(\hat{h})) \leq \frac{1}{n}D(\hat{\pi}\,\|\,\pi) + \frac{1}{n}\log\frac{2\sqrt{n}}{\delta}\right) \geq 1-\delta.$$

It remains to apply (3) and write

$$\mathsf{E}_{\hat{h}\sim\hat{\pi}}\delta(R_n(\hat{h})\,\|\,R(\hat{h})) \geq \delta\left(\mathsf{E}_{\hat{h}\sim\hat{\pi}}R_n(\hat{h})\,\middle\|\,\mathsf{E}_{\hat{h}\sim\hat{\pi}}R(\hat{h})\right).$$

References

[1] CATONI, O. *Statistical Learning Theory and Stochastic Optimization.* Springer, Berlin, 2004.

[2] LEDOUX, M., AND TALAGRAND, M. *Probability in Banach Spaces.* Springer-Verl., Berlin, 1991.

[3] MAURER, A. A note on the PAC Bayesian theorem. arXiv:cs/0411099v1.