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 $(x,y),(x_1,y_1),\ldots,(x_n,y_n)$ 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

$$\lambda(u) \coloneqq \begin{cases} 1 & \text{, if } u\leq 0, \\ 0 & \text{, otherwise,} \end{cases}$$

and the corresponding risk and empirical risk functions

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

Define the hinge loss

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

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

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

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

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

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

   $$\psi\circ\mathscr{A} \coloneqq \left\{ (\psi(a_1),\ldots,\psi(a_n)) \,\middle|\, (a_1,\ldots,a_n)\in\mathscr{A} \right\}.$$

   Then, it holds that

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

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

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

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

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

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

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

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.,

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

Problem 2

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

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

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

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

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

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

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:

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

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

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

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

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

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.**

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 $\varphi : \mathcal{H} \to \mathbb{R}$ possibly dependent on $z_1, \ldots, z_n$. The *change of measure inequality* says that for any $\eta \in]0, +\infty[$,

$$\eta \mathsf{E}_{\hat{h}\sim\hat{\pi}} \varphi(\hat{h}) \le D(\hat{\pi}\|\pi) + \log\left(\mathsf{E}_{\hat{h}\sim\pi} e^{\eta\varphi(\hat{h})} \right),$$

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 $\delta \in]0,1[$,**

$$\mathsf{P}\left(\mathsf{E}_{\hat{h}\sim\hat{\pi}}\varphi(\hat{h}) \leq \frac{1}{\eta}D(\hat{\pi}\|\pi) + \frac{1}{\eta}\log\frac{C_n(\eta)}{\delta}\right) \geq 1-\delta, \tag{2}$$

where the probability is with respect to the randomness of $z_1,\ldots,z_n$, and

$$C_n(\eta) \coloneqq \mathsf{E}_{z_1,\ldots,z_n}\mathsf{E}_{\hat{h}\sim\pi}\mathrm{e}^{\eta\varphi(\hat{h})}.$$

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

$$\mathsf{E}_{\hat{h}\sim\hat{\pi}}\varphi(\hat{h}) \coloneqq \sup_{B\in\mathbb{R}} \mathsf{E}_{\hat{h}\sim\hat{\pi}} \min\left\{B, \varphi(\hat{h})\right\}.$$

See [1, Section 5.2] for the details. *You can ignore this mathematical subtlety for this homework.*

2. (10 points) For any $p,q \in]0,1[$, define

$$\delta(p\|q) \coloneqq p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}.$$

Let u, v be random variables taking values in $]0,1[$. **Show that**

$$\mathsf{E}\,\delta(u\|v) \geq \delta(\mathsf{E}\,u\|\mathsf{E}\,v). \tag{3}$$

HINT: Let $\psi : \mathscr{X} \to \mathbb{R}$ for some set $\mathscr{X}$. *Jensen's inequality* says if for all $x, y \in \mathscr{X}$,

$$\psi((1-\alpha)x+\alpha y) \leq (1-\alpha)\psi(x) + \alpha\psi(y), \quad \forall\,\alpha\in[0,1],$$

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

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

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*

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

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 $n \geq 8$,**

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

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.