Theoretical assignment 2; 15 points total

Theoretical Deep Learning #2, MIPT

Let P be some prior over the set of predictors $\mathcal{F}$. Suppose we have a stochastic learning algorithm $\mathcal{A}$ which for every dataset S_n outputs a distribution $Q \,|\, S_n$ which we call a "posterior".

Let $\mathcal{D}$ be data distribution and $S_n = (x_i, y_i)_{i=1}^n \sim \mathcal{D}^n$ be training dataset. Let $\ell(y, f(x)) \in [0,1]$ be loss of predictor f on a pair (x, y).

Let $R(f) = \mathbb{E}_{(x,y)\sim\mathcal{D}} \ell(y, f(x))$ be expected risk of predictor f, and $\hat{R}_n(f) = \frac{1}{n} \sum_{(x,y)\in S_n} \ell(y, f(x))$ be training risk of predictor f. Let $R(Q) = \mathbb{E}_{f\sim Q} R(f)$ and $\hat{R}_n(Q) = \mathbb{E}_{f\sim Q} \hat{R}_n(f)$.

Given real numbers $q, p \in [0,1]$ define $KL(q \,\|\, p) = KL(\mathcal{B}(q) \,\|\, \mathcal{B}(p))$, where $\mathcal{B}(p)$ is a Bernoulli random variable with success probability p.

Problem 1

2.5 points total.

Let $p(x)$ and $q(x)$ be probability density functions (pdf's) defined on a set X.

1. **1.5 points.** Prove that for any function $h:\ X \to \mathbb{R}$

$$KL(p \,\|\, q) \geq \mathbb{E}_{x\sim p(x)} h(x) - \log \mathbb{E}_{x\sim q(x)} e^{h(x)}.$$

2. **1 point.** Prove that supremum over functions h is indeed a KL-divergence:

$$KL(p \,\|\, q) = \sup_{h:\ X\to\mathbb{R}} \left(\mathbb{E}_{x\sim p(x)} h(x) - \log \mathbb{E}_{x\sim q(x)} e^{h(x)} \right).$$

Problem 2

2 points.

Assume there exists a dataset negation procedure "$\neg(\cdot)$" with following properties:

1. $\neg(\neg(S_n)) = S_n \quad \forall S_n$;
2. $KL(Q \,|\, S_n \,\|\, P)) = KL(Q \,|\, \neg(S_n) \,\|\, P)) \quad \forall S_n$;
3. $\hat{R}_n(Q \,|\, S_n) = 0 \quad \forall S_n$;

4. $\hat{R}_n(Q \mid \neg(S_n)) = 1 \quad \forall S_n$;
5. $R(Q \mid S_n) < \epsilon \quad \forall S_n$.

Prove that

$$\sqrt{\frac{1}{2n-1}\left(\log\frac{4n}{\delta} + KL(Q \mid S_n \parallel P)\right)} \geq 1 - \epsilon \quad \forall S_n.$$

From this follows that if above-defined dataset negation procedure exists, PAC-bayesian bound of McAllester (1999)[1] *becomes nearly-vacuous.*

Problem 3

2 points.
Consider a PAC-bayesian bound in the form of Langford & Seeger (2001)[2]:

$$KL(\hat{R}_n(Q \mid S_n) \parallel R(Q \mid S_n)) \leq \frac{1}{n-1}\left(\log\frac{2n}{\delta} + KL(Q \mid S_n \parallel P)\right)$$

w.p. $\geq 1 - \delta$ over S_n.

Let the stochastic learning algorithm $\mathcal{A}$ which produces "posterior" distributions $Q \mid S_n$ be given. What will be the optimal prior distribution? More concretely, find P which minimizes right-hand side expected over training datasets:

$$\text{Find } P \in \operatorname*{Arg\,min}_P \mathbb{E}_{S_n}\left(\log\frac{2n}{\delta} + KL(Q \mid S_n \parallel P)\right).$$

Problem 4

4 points.
Consider a PAC-bayesian bound in the form:

$$R(Q \mid S_n) \leq 2\left(\hat{R}_n(Q \mid S_n) + \frac{1}{n}\left(\log\frac{1}{\delta} + KL(Q \mid S_n \parallel P)\right)\right)$$

w.p. $\geq 1 - \delta$ over S_n.

Let prior P and training dataset S_n be given. What will be the optimal "posterior" distribution? More concretely, find Q which minimizes right-hand side:

$$\text{Find } Q \in \operatorname*{Arg\,min}_Q \left(\hat{R}_n(Q) + \frac{1}{n}\left(\log\frac{1}{\delta} + KL(Q \parallel P)\right)\right).$$

Here assume that both prior and "posterior" distributions have densities.

[1] Theorem 2 in `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.21.1908&rep=rep1&type=pdf`

[2] Theorem 3 in `http://hunch.net/~jl/projects/prediction_bounds/averaging/averaging_tech.pdf`

Problem 5

3 points.

Consider a more general PAC-bayesian bound:

$$R(Q \mid S_n) \leq \frac{1}{1 - \frac{1}{2\lambda}} \left(\hat{R}_n(Q \mid S_n) + \frac{\lambda}{n} \left(\log \frac{1}{\delta} + KL(Q \mid S_n \parallel P) \right) \right)$$

w.p. $\geq 1 - \delta$ over S_n $\forall \lambda > 1/2$.

Suppose the space of predictors $\mathcal{F}$ be finite. Again, let prior P and training dataset S_n be given. Find Q which minimizes right-hand side:

$$\text{Find } Q \in \operatorname*{Arg\,min}_Q \left(\hat{R}_n(Q) + \frac{\lambda}{n} \left(\log \frac{1}{\delta} + KL(Q \parallel P) \right) \right).$$

Problem 6

1.5 points total.

Consider a PAC-bayesian bound similar to the bound of Langford & Seeger (2001):

$$KL_\gamma(\hat{R}_n(Q \mid S_n) \parallel R(Q \mid S_n)) \leq \frac{1}{n} \left(\log \frac{1}{\delta} + KL(Q \mid S_n \parallel P) \right)$$

w.p. $\geq 1 - \delta$ over S_n $\forall \gamma \in \mathbb{R}$, where γ-KL-divergence between real numbers $q, p \in [0, 1]$ is defined as follows:

$$KL_\gamma(q \parallel p) = \gamma q - \log(1 - p + pe^\gamma).$$

1. **0.5 points.** Prove that $\sup_\gamma KL_\gamma(q \parallel p) = KL(q \parallel p)$;
2. **1 point.** Given previous statement, does the bound above imply the following bound?:

$$KL(\hat{R}_n(Q \mid S_n) \parallel R(Q \mid S_n)) \leq \frac{1}{n} \left(\log \frac{1}{\delta} + KL(Q \mid S_n \parallel P) \right)$$

   w.p. $\geq 1 - \delta$ over S_n. Argue, why.