TTIC 31230 Fundamentals of Deep Learning, winter 2020

Quiz 3

PAC-Bayes Background for the problem 1. Consider any probability distribution $P(h)$ over a discrete class $\mathcal{H}$. Assume $0 \leq \mathcal{L}(h,x,y) \leq L_{\max}$. Define

$$\mathcal{L}(h) = E_{(x,y)\sim \text{Pop}}\, \mathcal{L}(h,x,y)$$

$$\hat{\mathcal{L}}(h) = E_{(x,y)\sim \text{Train}}\, \mathcal{L}(h,x,y)$$

We now have the theorem that with probability at least $1-\delta$ over the draw of training data the following holds simultaneously for all h.

$$\mathcal{L}(h) \leq \frac{10}{9}\left(\hat{\mathcal{L}}(h) + \frac{5L_{\max}}{N}\left(\ln\frac{1}{P(h)} + \ln\frac{1}{\delta}\right)\right) \quad (1)$$

This motivates

$$h^* = \operatorname*{argmin}_h\ \hat{\mathcal{L}}(h) + \frac{5L_{\max}}{N}\ln\frac{1}{P(h)} \quad (2)$$

The Bayesian maximum a-posteriori (MAP) rule is

$$h^* = \operatorname*{argmax}_h\ P(h)\prod_{(x,y)\in\text{Train}} P(y|x,h) \quad (3)$$

Problem 1. Finite Precision Parameters. (25 points)

(a) Consider a model where the parameter vector Φ has d parameters each of which is represented by a 16 bit floating point number. Express the bound (1) in terms of the dimension d assuming all parameter vectors are equally likely.

Solution:

$$\mathcal{L}(h) \leq \frac{10}{9}\left(\hat{\mathcal{L}}(h) + \frac{5L_{\max}}{N}\left(16d\ln 2 + \ln\frac{1}{\delta}\right)\right)$$

(b) Assume a variable precision representation of numbers where $\Phi[i]$ is given with $|\Phi[i]|$ bits. Express the bound (1) as a function of Φ assuming that $P(\Phi)$ is defined so that each parameter is selected independently and that

$$P(\Phi[i]) = 2^{-|\Phi[i]|}$$

.

Solution:

$$\mathcal{L}(h) \leq \frac{10}{9}\left(\hat{\mathcal{L}}(h) + \frac{5L_{\max}}{N}\left(|\Phi|\ln 2 + \ln\frac{1}{\delta}\right)\right)$$

$$|\Phi| = \sum_i |\Phi[i]|$$

(c) Repeat part (a) but for a model with d parameters represented by $\Phi_i = z[J[i]]$ where $J[i]$ is an integer index with $0 \leq J[i] < k$ and where $z[j]$ is a b bit floating point number and where all parameter vectors are equally likely.

Solution:

$$\mathcal{L}(h) \leq \frac{10}{9}\left(\hat{\mathcal{L}}(h) + \frac{5L_{\max}}{N}\left(kb\ln 2 + d\ln k + \ln\frac{1}{\delta}\right)\right)$$

Since d is large this is typically much tighter bound than using floating point or even integer representationfs of parameters. It is a much more compact representaiton of the parameters.

Problem 2. Pseudolikelihood of a one dimensional spin glass. (25 points) We let $\hat{x}$ be an assignment of a value to every node where the nodes are numbered from 1 to N_{nodes} and for every node i we have $\hat{x}[i] \in \{0, 1\}$. We define the score of $\hat{x}$ by

$$f(\hat{x}) = \sum_{i=1}^{N-1} \mathbf{1}[\hat{x}[i] = \hat{x}[i+1]]$$

The probability distribution over assignments is defined by a softmax. We let $\hat{x}[i := v]$ be the assignment identical to $\hat{x}$ except that node i is assigned the value v. The expression $\hat{x}[i] = v$ is either true or false depending on whether no i is assigned value v in $\hat{x}$. So these are quite different.

$$P_f(\hat{x}) = \operatorname*{softmax}_{\hat{x}} f(\hat{x})$$

Pseudolikelihood is defined in terms of the softmax probability P_f as follows.

$$\tilde{P}_f(\hat{x}) \quad = \quad \prod_i P_f(\hat{x}[i] \mid \hat{x}\backslash i)$$

What is the pseudolikelihood of the all ones assignment under the definition of f given above?

Solution: In a graphical model $P_f(\hat{x}[i] \mid \hat{x}/i)$ is determined by the neighbors of i and we can consider only how a value is scored against it neighbors. For $\hat{x}$ equal to all ones we have

$$f(\hat{x}) \quad = \quad N - 1$$

$$f(\hat{x}[i := 0]) \quad = \quad \begin{cases} N-3 & \text{for } 1 < i < N \\ N-2 & \text{for } i = 1 \text{ or } i = N \end{cases}$$

For $1 < i < N$ we get

$$\begin{aligned} Q_f(\hat{x}[i=1] \mid \hat{x}/i) &= \frac{e^{N-1}}{e^{N-1} + e^{N-3}} \\ &= \frac{1}{1 + e^{-2}} \end{aligned}$$

and for $i = 1$ or $i = N$ we get

$$Q_f(\hat{x}[i=1] \mid \hat{x}/i) = \frac{1}{1 + e^{-1}}$$

This gives

$$\tilde{Q}(\hat{x}) = (1 + e^{-1})^{-2}(1 + e^{-2})^{-(N-2)}$$

Problem 3. Pseudolikelihood for Monocular Distance Estimation. (25 points) Here we are interested in labeling each pixed with a distance from the camera. Each pixel i is to be labeled with a real number $\hat{y}[i] > 0$ giving the distance in (say) meters from the camera to the point on the object displayed by that pixel. We assume a neural network that computes for each pixel i an expected distance μ_i and a variance $\sigma_i > 0$. For each pair of neighboring pixels i and j the neural network computes a real number $\lambda_{\langle i,j \rangle} \geq 0$. For each assignment $\hat{y}$ of distances to pixels we then define the score $s(\hat{y})$ by

$$s(\hat{y}) = \sum_{i \in \text{nodes}} -(\hat{y}[i] - \mu_i)^2/\sigma_i^2 + \sum_{\langle i,j \rangle \in \text{edges}} -\lambda_{\langle i,j \rangle} |\hat{y}[i] - \hat{y}[j]|$$

(a) This scoring function determines a continuous softmax distribution defined by

$$p(\hat{y}) = \frac{1}{Z} e^{s(\hat{y})}$$

where Z is an integral rather than a sum. What is the dimension of the space to be integrated over in computing Z?

Solution: This is an integration over $\mathbb{R}^N$ where N is the number of nodes — an N_{nodes} dimentional space.

(b) We now consider pseudolikelihood for this problem. Give an expression for the continuous conditional probability density on $\hat{y}[i]$ for the distance $\hat{y}[i]$ conditioned on the value of the neighbors $N(i)$ of node i. This probability is written $p(\hat{y}[i] \mid \hat{y}[N(i)])$. You answer should be given as a function of the values $\hat{y}[j]$ for the nodes j neighboring i written $j \in N(i)$. Write Z as an integral but do not bother trying to solve it. What is the dimention of the integral for this conditional probability?

Solution:

$$p(\hat{y}[i] \mid \hat{y}[N(i)]) = \frac{1}{Z} \exp\left(-(\hat{y}[i] - \mu_i)^2/\sigma_i^2 + \sum_{j \in N(i)} -\lambda_{\langle i,\, j\rangle} |\hat{y}[i] - \hat{y}[j]| \right)$$

$$Z = \int_0^\infty \exp\left(-(x - \mu_i)^2/\sigma_i^2 + \sum_{j \in N(i)} -\lambda_{\langle i,\, j\rangle} |x - \hat{y}[j]| \right) dx$$

This is an integral over a one dimensional space (a single real number).

Problem 4. Generalization Bounds for the realizable case. (25 points) Consider a finite hyothesis class $\mathcal{H}$ and a population distribution Pop on pairs $\langle x, y\rangle$ such that for $\langle x, y\rangle$ drawn from the population and $h \in \mathcal{H}$ we have that h makes a prediction for y which we will write as $h(x)$. The error rate of hypothesis h on the population is defined by

$$\text{Err}_{\text{Pop}}(h) = P_{\langle x,\, y\rangle \sim \text{Pop}}(h(x) \neq y)$$

We draw a training sample Train consisting of N_{Train} pairs $\langle x, y\rangle$ drawn IID from the population.

$$\text{Err}_{\text{train}}(h) = \frac{1}{N_{\text{train}}} \sum_{\langle x,\, y\rangle \in \text{Train}} \mathbf{1}(h(x) \neq y)$$

(a) For a given hypothesis h with error rate ϵ what is the probability that $\text{Err}_{\text{train}}(h) = 0$.

Solution: $(1 - \epsilon)^{N_{\text{train}}}$

(b) We now consider a fixed threshold ϵ and consider the hypotheses h satisfying $\text{Err}_{\text{Pop}} \geq \epsilon$. We will call these the "bad" hypotheses.

The simple form of the union bound is

$$P(A \cup B) \leq P(A) + P(B)$$

This can be generalized to

$$P(\exists z\ Q(z)) \leq \sum_z P(Q(z))$$

where $Q(z)$ is any statement about z.

Use your answer to (a) and the union bound to give an upper bound on the probability that there exists a bad hypothesis h with $\text{Err}_{\text{train}}(h) = 0$. You solution should be stated in terms of ϵ, the number of elements $|\mathcal{H}|$ of $\mathcal{H}$, and

the number of training pairs N_{train}. Simplify your solution using the inequality $1 - \epsilon \leq e^{-\epsilon}$.

Solution:

$$
\begin{aligned}
& P(\exists h \ \ \text{Err}_{\text{Pop}} \geq \epsilon, \ \text{Err}_{\text{train}}(h) = 0) \\
\leq & \sum_{h:\ \text{Err}_{\text{Pop}}(h) \geq \epsilon} P(\text{Err}_{\text{train}}(h) = 0) \\
\leq & |\mathcal{H}|(1-\epsilon)^{N_{\text{train}}} \\
\leq & |\mathcal{H}| e^{-N_{\text{train}}\epsilon}
\end{aligned}
$$

(c) Now consider a small positive number δ and solve for ϵ such that the probability that a bad hypothesis has zero training error is less than δ. Your solution gives a value of ϵ such that with probability $1 - \delta$ over the draw of the training error all hypothesis with zero training error have population error no lareger than ϵ.

Solution:

$$
\begin{aligned}
\delta &= |\mathcal{H}| e^{-N_{\text{train}}\epsilon} \\
\epsilon &= \frac{\ln |\mathcal{H}| + \ln \frac{1}{\delta}}{N_{\text{train}}}
\end{aligned}
$$