Advancd Topics in Machine Learning - Assignment 2

Christoffer Thrysøe - dfv107

September 18, 2017

1. Subgradients

1

We wish to construct a function $f : \mathcal{R} \to \mathcal{R}$ with differential set $[-2, 2]$ at 0.
Choosing the function

$$f(x) = 2|x| \tag{1}$$

we have that for $x < 0$ the subgradient is $\partial f(x) = -2$ and for $x > 0$ we have $\partial f(x) = 2$ therefore at 0, the differential set for f is $[-2, 2]$.

2

For the function to have the desired properties we must have:

$$f(z) \geq f(x) + g^T(z - x) \tag{2}$$

where g is the differential set, and $x = 0$, thus we have the following:

$$2|z| \geq gz \tag{3}$$

clearly the desired property is satisfied when $g \in [-2, 2]$ because

$$2|z| \geq 2z \tag{4}$$

$$2|z| \geq -2z \tag{5}$$

2. Convexity of logistic regression

1

We wish to prove that the derivative of

$$f(x) = \ln(1 + e^x) \tag{6}$$

is

$$f'(x) = \frac{1}{1 + e^{-x}} \tag{7}$$

For the proof we first note that the derivative of $\ln(x)$ is $1/x$, then we apply the chain rule as followed:

$$\frac{\partial}{\partial x} f(x) = \frac{\partial}{\partial x} \left[ln(1 + e^x) \right] \tag{8}$$

$$= \frac{1}{1 + e^x} \frac{\partial}{\partial x} \left[e^x \right] \tag{9}$$

$$= \frac{e^x}{1 + e^x} \tag{10}$$

To get (7) I will instead start at (7) and derive (10).
First we note that $e^{-x} = 1/e^x$ and we get the following:

$$\begin{aligned} \frac{1}{1 + e^{-x}} &= \frac{1}{\frac{1}{1} + \frac{1}{e^x}} \\ &= \frac{1}{\frac{e^x}{e^x} + \frac{1}{e^x}} \\ &= \frac{1}{\frac{1+e^x}{e^x}} \\ &= \frac{e^x}{1 + e^x} \end{aligned}$$

and therefore we can conclude that:

$$f'(x) = \frac{1}{1 + e^{-x}}$$

To prove that the function $f'(x)$ is a monotonically increasing function, we must prove that $f''(x)$ is positive. First we take the derivative of (7):

$$\frac{\partial}{\partial x} f'(x) = \frac{\partial}{\partial x} \left[\frac{1}{1 + e^{-x}} \right]$$

Again we apply the chain rule where we have:

$$\frac{\partial y}{\partial x} = \frac{\partial y}{\partial u} \frac{\partial u}{\partial x}$$

in which we define $u = 1 + e^{-x}$, $y = u^{-1}$, thus we have:

$$\begin{aligned} \frac{\partial y}{\partial u} &= -\frac{1}{u^2} \\ \frac{\partial u}{\partial x} &= -e^{-x} \end{aligned}$$

and combined

$$\begin{aligned} \frac{\partial y}{\partial x} &= -\frac{1}{(1 + e^{-x})^2} - e^{-x} \\ &= -\frac{-e^{-x}}{(1 + e^{-x})^2} \\ &= \frac{e^{-x}}{(1 + e^{-x})^2} \end{aligned}$$

and therefore:

$$f''(x) = \frac{e^{-x}}{(1 + e^{-x})^2} \tag{11}$$

To see why (11) is positive, we note that e^{-x} converges to zero but is positive for all $x > 0$, therefore we have that $f'(x) > 0$ and is thus a monotonically increasing function.

3. PAC-Bayes vs. Occam

1

We wish to prove that with probability greater than $1-\delta$:

$$\mathrm{kl}\left(\mathbb{E}_\rho\left[\hat{L}(h,S)\right] \parallel \mathbb{E}_\rho\left[L(h)\right]\right) \leq \frac{\sum\limits_{h\in\mathcal{H}} \rho(h)\ln\frac{1}{\pi(h)} + \ln\frac{n+1}{\delta}}{n} \tag{12}$$

First we note that:

$$\mathrm{KL}(\rho||\pi) = \mathbb{E}_\rho\left[\ln\frac{1}{\pi}\right] - H(\rho) \tag{13}$$

where the entropy $H(p)$ is defined as $H(p) = -\sum\limits_{x\in\mathcal{X}} p(x)\,\ln\,p(x)$ thus $H(p)$ is negative. From theorem 3.13 from Yevgeny's lecture notes, we have the following:

$$P\left\{\mathrm{kl}\left(\mathbb{E}_\rho\left[\hat{L}(h,S)\right] \parallel \mathbb{E}_\rho\left[L(h)\right]\right) \geq \frac{\mathrm{KL}(\rho||\pi) + \ln\frac{n+1}{\delta}}{n}\right\} \leq \delta \tag{14}$$

if we input the definition of KL from (13) in (14), we get the following

$$P\left\{\mathrm{kl}\left(\mathbb{E}_\rho\left[\hat{L}(h,S)\right] \parallel \mathbb{E}_\rho\left[L(h)\right]\right) \geq \frac{\sum\limits_{h\in\mathcal{H}} \rho(h)\ln\frac{1}{\pi(h)} - H(\rho(h)) + \ln\frac{n+1}{\delta}}{n}\right\} \leq \delta \tag{15}$$

Because $-H(\rho)$ is positive, we can remove it from (15) and the bound will still hold. Thus we can write for all distributions of ρ over $\mathcal{H}$ we have the following bound:

$$P\left\{\mathrm{kl}\left(\mathbb{E}_\rho\left[\hat{L}(h,S)\right] \parallel \mathbb{E}_\rho\left[L(h)\right]\right) \leq \frac{\sum\limits_{h\in\mathcal{H}} \rho(h)\ln\frac{1}{\pi(h)} + \ln\frac{n+1}{\delta}}{n}\right\} \geq 1-\delta \tag{16}$$

which is what we wanted to prove.

2

We wish to prove that the PAC-bayes-kl inequality, which holds with probability greater than $1-\delta$:

$$\mathrm{kl}\left(\mathbb{E}_\rho\left[\hat{L}(h,S)\right] \parallel \mathbb{E}_\rho\left[L(h)\right]\right) \leq \frac{\mathrm{KL}(\rho||\pi) + \ln\frac{n+1}{\delta}}{n} \tag{17}$$

is atleast as tight as the Occam's razor bound with kl, which holds with probability greater than $1-\delta$:

$$\mathrm{kl}\left(\mathbb{E}_\rho\left[\hat{L}(h,S)\right] \parallel \mathbb{E}_\rho\left[L(h)\right]\right) \leq \frac{\sum\limits_{h\in\mathcal{H}} \rho(h)\ln\frac{1}{\pi(h)} + \ln\frac{n+1}{\delta}}{n} \tag{18}$$

Thus we wish to show that:

$$\frac{\sum\limits_{h\in\mathcal{H}} \rho(h)\ln\frac{1}{\pi(h)} + \ln\frac{n+1}{\delta}}{n} \leq \frac{\mathrm{KL}(\rho||\pi) + \ln\frac{n+1}{\delta}}{n} \tag{19}$$

If we again take the definition of KL, we get the following:

$$\frac{\mathrm{KL}(\rho||\pi) + \ln\frac{n+1}{\delta}}{n} = \frac{\sum_{h\in\mathcal{H}} \rho(h)\ln\frac{1}{\pi(h)} - H(\rho(h)) + \ln\frac{n+1}{\delta}}{n} \tag{20}$$

$$\geq \frac{\sum_{h\in\mathcal{H}} \rho(h)\ln\frac{1}{\pi(h)} + \ln\frac{n+1}{\delta}}{n} \tag{21}$$

where the last inequality follows from $-H(\rho)$ being positive, thus proving that the PAC-bayes inequality is atleast as tight as the Occam's razor bound with kl.

4. Nonnegativity of KL

1

We wish to prove that for all $0 < x < \infty$:

$$\ln(x) \leq x - 1 \tag{22}$$

To show this, we construct a function of the left hand side minus the right hand side:

$$f(x) = \ln(x) - x + 1 \tag{23}$$

If we can prove that $f(x) \leq 0$ for all $x > 0$ we will have proven (22). The strategy of doing so is to locate the functions local maximum and prove that this is a global maximum, and is smaller than 0. First we take the derivative of (23):

$$\frac{\partial}{\partial x} f(x) = \frac{\partial}{\partial x}\left[\ln(x) - x + 1\right] \tag{24}$$

$$= \frac{1}{x} - 1 \tag{25}$$

Finding the optimum:

$$f'(x) = \frac{1}{x} - 1 = 0 \Leftrightarrow x = 1 \tag{26}$$

The function value of the optimum is $f(1) = \ln(1) - 1 + 1 = 0$. To see if the optimum is a local maximum we take the second derivative of f:

$$\frac{\partial^2}{\partial x^2} f(x) = \frac{\partial^2}{\partial x^2}\left[\ln(x) - x + 1\right] \tag{27}$$

$$= -\frac{1}{x^2} \tag{28}$$

since we have that $f''(1) = -\frac{1}{1^2} = -1$ we have that $f(1)$ is a local maximum and that the function increases from 0 to 1 until a stationary point is met at $x = 1$ and then the function decreases afterwards, thus being concave in $x = 1$. We now need to determine that $f'(x) < 0$ for all $x > 1$. This can easily be verified by looking at $f'(x)$ shown in (25), as x increases $1/x$ will decrease and therefore we have that $f(x) < 0\ \forall x > 1$. Which means that $f(0)$ is global maximum and we have that:

$$\forall x > 0 < \infty : f(x) \leq f(1)$$

which verifies that:

$$ln(x) \leq x - 1$$

2

We wish to prove that

$$KL(p||q) \geq 0 \tag{29}$$

First we note that the definition of $KL(p||q)$ is given by:

$$KL(p||q) = \sum_{x \in X} p(x) \ln \frac{p(x)}{q(x)} \tag{30}$$

(we assume that p and q are discrete distributions as the assignment text allows). For the proof we wish to show that $-KL(p||q) \leq 0$:

$$\begin{aligned}
-KL(p||q) &= -\sum_{x \in X} p(x) \ln \frac{p(x)}{q(x)} && (31)\\
&= \sum_{x \in X} p(x) \ln \frac{q(x)}{p(x)} && (32)\\
&\leq \sum_{x \in X} p(x) \left(\frac{q(x)}{p(x)} - 1 \right) && (33)\\
&= \sum_{x \in X} p(x) \left(\frac{q(x) - p(x)}{p(x)} \right) && (34)\\
&= \sum_{x \in X} p(x) - q(x) && (35)\\
&= 0 && (36)
\end{aligned}$$

where (32) follows from: $-a \ln (b/c) = a \ln (c/b)$, (33) follows from the previous proof that $\ln (x) \leq x - 1$, (34) follows from $(a/b) - 1 = (a - b)/b$ and (36) follows from $\sum_{x \in X} p(x) = 1$ and $\sum_{x \in X} q(x) = 1$

which verifies that $KL(p||q) \geq 0$

If we have that $p(x) = 0$ for any x, the inequality on line (33), does not hold because $ln\frac{0}{q} = 0$, thus we end up with $kl(p||q) = 0$. If we on the other hand have that for some x: $q(x) = 0$, then $p(x)\ln\frac{p}{0} = \infty \geq 0$, thus we have proved that $kl(p||q) \geq 0$

3

If we have that $p(x) = q(x)$, the condition $\ln (x) \leq x - 1$ does not hold, as we have the following on line (33) in the proof:

$$\begin{aligned}
\ln (x) \leq x - 1 \Rightarrow && (37)\\
\ln \frac{q(x)}{p(x)} \leq \frac{q(x)}{p(x)} - 1 && (38)
\end{aligned}$$

however because we have that $p(x) = q(x)$ we have the following:

$$\begin{aligned}
\ln \frac{q(x)}{p(x)} = \frac{q(x)}{p(x)} - 1 \Rightarrow && (39)\\
\ln (1) = \frac{1}{1} - 1 \Rightarrow && (40)\\
0 = 0 && (41)
\end{aligned}$$

Thus we end up with an equality instead of an inequality and get $Kl(p||q) = 0$ when $p(x) = q(x)$.

5. Early stopping and overfitting

1

For this question, I will first discuss the usage of the validation set for early stopping. For $\hat{L}(h_{t^*}, S^{val})$ to be an unbiased estimate of $L(h_{t^*})$ we must have that $\mathbb{E}[\hat{L}(h_{t^*}, S^{val})] = L(h_{t^*})$. Since we have not used the validation set S^{val} for training, we can use the validation set to form an unbiased estimate of the expected error $L(h_{t^*})$. When we, however choose our final hypothesis, based on the validation set S^{val}, we introduce a small bias, because we are somewhat basing the training process (model selection) on the validation set. The error on the validation set will fluctuate, unlike the test set, and thus it will have a variance. The error can have an "optimistic" variance i.e. the error is low or a "pessimistic" variance in which the error is high, these sudden fluctuations may be coincidental and not reflect the test error. If we take the lowest error on the validation set to represent the error on the test set, we are being optimistic also about the test set, and thus we introduce an optimistic bias in the validation error. To show this bias, we consider the following example:
Given two hypothesis h_1, h_2 with $L(h_1) = L(h_2) = 0.5$, where the actual error l_1, l_2 are uniformly selected in the interval $[0, 1]$. Each error is an unbiased estimate of the expected loss. If we choose the smallest error of l_1, l_2 (similar as choosing the smallest validation error), i.e. we take: $l = \min(l_1, l_2)$. Now, however we have that the expected value is less than 0.5 with probability of 75%. This is because we are picking the smallest loss, and we only need one to be lower than 0.5, which it is with probability of 75%. Due to the selection of models the errors are no longer unbiased estimates of their expected loss, due to the optimistic bias. This example shows that choosing the model based on the validation error introduces a small bias.

a

For this approach, we stop the training after 100 iterations and return $h_t^* = h_{100}$. Because we are not stopping based on the validation error, but simply on a given number of training iteration the error $\hat{L}(h_{t^*}, S^{val})$ is an unbiased estimate of $L(h_{t^*})$, because S^{val} did not have any influence on the model selection.

b

For this approach, we stop the training once we observe a training iteration where $\hat{L}(h_i, S^{val}) > \hat{L}(h_{i-1}, S^{val})$, then $t^* = i^* - 1$. The model selection is influenced by the validation set. Therefore $\hat{L}(h_{t^*}, S^{val})$ is not an unbiased estimate of $L(h_{t^*})$.

c

For this approach, we stop the training if we observe a given number of training iterations, which did not result in a decrease in the validation error, and return the best model observed so far. Again the model selection is influenced by the validation set, and introduces more optimistic bias than the previous early stopping method, because it looks for the smallest validation error over even more training cycles. Therefore $\hat{L}(h_{t^*}, S^{val})$ is not an unbiased estimate of $L(h_{t^*})$.

2

For approach a, we only consider a single hypothesis, i.e. when $t = 100$, thus we can use Hoeffding's to get a bound on $L(h)$, which holds with probability greater than $1 - \delta$:

$$L(h_{t^*}) \leq \hat{L}(h_{t^*}, S^{val}) + \sqrt{\frac{\ln \frac{1}{\delta}}{2n}} \tag{42}$$

where $t^* = 100$.
For b and c we note that choosing the model h_{t^*}, we have seen t^* models. As more models are being considered, the probability of encountering validation errors which got "lucky" and will not represent the expected error increases. Thus, the more hypothesis we consider, the looser the bound should be. We have a finite number of hypothesis at t^*, and we wish to have an increasing bound as the number of hypothesis increases. Therefore we can use Occam's Razor bound and define $p(h_i)$ to increase the complexity term based on the number of hypothesis. We can use the following function of $p(h_i)$ to get such a behaviour:

$$p(h_i) = \frac{1}{2^i} \tag{43}$$

Thus we can derive the following bound on on $L(h_{t^*})$, which holds with probability greater than $1 - \delta$:

$$L(h_{t^*}) \leq \hat{L}(h_{t^*}, S^{val}) + \sqrt{\frac{\ln \frac{1}{p(h_t^*)\delta}}{2n}} \tag{44}$$

where t^* is the number of hypothesis considered at learning epoch t^* as each training epoch gives us a new model. We could also have used Occam's razor bound, and defined $p(h)$ to loosen the bound as more hypothesis are used, but as we are only bounding the probability for h_{t^*}, we can use the generalization bound for finite hypothesis.

3

b

As the bound from question 2 indicates, the larger the number of considered hypothesis gets, the less trust we have in the generalization. Thus we could add an increasing penalty on the validation error as t increases. We could add the term of the bound in (44), that is for training epoch t^i, we redifine the validation error:

$$\hat{L}(h_{t^i}, S^{val}) = \hat{L}(h_{t^i}, S^{val}) + \sqrt{\frac{\ln \frac{t^i}{\delta}}{2n}} \tag{45}$$

and again stop, whenever we observe an increase in the validation error, thus when:

$$\hat{L}(h_{t^i}, S^{val}) + \sqrt{\frac{\ln \frac{t^i}{\delta}}{2n}} < \hat{L}(h_{t^{i+1}}, S^{val}) + \sqrt{\frac{\ln \frac{t^{i+1}}{\delta}}{2n}} \tag{46}$$

However as discussed, the validation error is a fluctuating function, therefore the first decrease in $\hat{L}(h_{t^i}, S^{val})$ is likely to be a random fluctuation, instead of indicating a good stopping point, thus I would use a similar approach to c.

c

We can use the same approach of penalizing the validation error as the number of hypothesis grow by adding the last term of the bound in (44). The early stopping approach would remain the same, thus we are still going x epochs after finding an increase in the validation error, but with the added penalty for considering more hypothesis. Thus after encountering an increase in the validation error + bound (as shown in (46)) , we check the following:

$$\hat{L}(h_{t^{opt}}, S^{val}) + \sqrt{\frac{\ln \frac{t^{opt}}{\delta}}{2n}} < \hat{L}(h_{t^{i+1}}, S^{val}) + \sqrt{\frac{\ln \frac{t^{i+1}}{\delta}}{2n}} \tag{47}$$

where t^{opt} is the training epoch, which resulted in the lowest achieved validation error + bound, before we encountered an increase in the validation error. If (47) does not hold, we have found a better validation error, and will use this as t^{opt} . Afer x iterations have passed the optimal value will be returned.

4

We wish to use the bound to define, when we can definently stop, after observing an increase in the validation error, that is when the bound will definetely stop improving.