Machine Learning - Assignment 4

Christoffer Thrysøe - dfv107

December 19, 2016

Finite Hypothesis Space

Question 1

Below is the hypothesis space in the first and second approach:

1. In the first approach every possible individual combination of the pair: $\{age, gender\}$ are considered, where $gender \in \{\text{male,female}\}$ & $age \in \{0, ..., 100\}$. Each hyptohesis has a binary outcome, therefore the size of above space is:

$$|\{0,1\}|^{|\{0,..,100\}| \times |\{\text{male,female}\}|} = 2^{101 \times 2} = 2^{202} = 6.43 \times 10^{60}$$

2. For the second approach, for each tuple (i, j) we chose the set of indices i,j such that : $0 \leq i < j \leq 100$. The indices are chosen from 101 possible age values thus we have the set $\binom{101}{2} = 5050$ and for both gender we have the hypothesis size:

$$\binom{101}{2}^2 = 5050^2 = 25502500$$

Question 2

To write a high probability bound on $L(h)$ in terms of $\hat{L}(h, S)$ we use the following bound:

$$Pr\left[\exists h \in \mathcal{H} : L(h) \geq \hat{L}(h, S) + \sqrt{\frac{ln\frac{M}{\delta}}{2n}}\right] \leq \delta$$

Which we can write the complement of for all $h \in \mathcal{H}$:

$$Pr\left[\forall h \in \mathcal{H} : L(h) \leq \hat{L}(h, S) + \sqrt{\frac{ln\frac{M}{\delta}}{2n}}\right] \geq 1 - \delta \qquad (1)$$

Given the bound for $L(h)$ from equation 1, we can plug in the numbers for the hypothesis space for the two cases:

1.
$$Pr\left[\forall h \in \mathcal{H} : L(h) \leq \hat{L}(h,S) + \sqrt{\frac{ln\frac{6.43\cdot 10^{60}}{\delta}}{2n}}\right] \geq 1-\delta$$

The complexity is written as $p(h) = \frac{1}{M} = \frac{1}{6.43\cdot 10^{60}}$

2.
$$Pr\left[latexr\forall h \in \mathcal{H} : L(h) \leq \hat{L}(h,S) + \sqrt{\frac{ln\frac{25502500}{\delta}}{2n}}\right] \geq 1-\delta$$

The complexity is written as $p(h) = \frac{1}{M} = \frac{1}{25502500}$

Question 3

Choosing a limited hypothesis set results in a bigger chance of the empirical loss being equal to the actual loss. However a reduced hypothesis set may also lead to under-fitting of the data, as the set of hypothesis may not represent the true learning algorithm. Likewise a not restrictive hypothesis may lead to over-fitting. Choosing a range as the hypothesis is not advantageous when dealing with distribution not entered around a mean, for example a single range of age doesn't describe when people are most likely to visit the dentist.

Occam's Razor

Question 1

First we note that the size of the alphabet is $|\sum| = 27$, given a string of d characters, the possible space must be: 27^d, for each of these mappings is an output $|\{0,1\}|$ therefore we have the hypothesis space (possible mappings): 2^{27^d}. Now we can use equation 1 again to derive at high probability bound on $L(h)$ for all $h \in \mathcal{H}$:

$$Pr\left[\forall h \in \mathcal{H} : L(h) \leq \hat{L}(h,S) + \sqrt{\frac{ln\frac{2^{27^d}}{\delta}}{2n}}\right] \geq 1-\delta$$

Question 2

We start by setting $p(h) = 1/2^{d(h)}2^{27^d}$ we therefore have that:

$$\sum_{h\in\mathbb{H}} p(h) = \sum_{d=0}^{\infty}\sum_{h\in\mathcal{H}_d} 1/2^d 2^{27^d} = \sum_{d=0}^{\infty} |\mathcal{H}_d| 1/2^d 2^{27^d} = \sum_{d=0}^{\infty} 1/2^d = 1$$

we can apply Occam's Razor and obtain a high probability bound for all $h \in \mathcal{H}$:

$$Pr\left[\forall h \in \mathcal{H} : L(h) \leq \hat{L}(h,S) + \sqrt{\frac{ln\frac{2^{d(h)}2^{27^{d(h)}}}{\delta}}{2n}}\right] \geq 1-\delta$$

Question 3

The advantages of choosing a large d is the possibility to describe the data, with more hypothesis, thus avoiding under-fitting. However choosing a large d may also lead to over-fitting of the data.

3. Logistic regression

3.1 Cross-entropy error measure

a

We have the likelihood function:

$$Pr\,\{\,y\,|\,\mathbf{x}\,\} = \begin{cases} h(\mathbf{x}) & \text{for } y = +1 \\ 1 - h(\mathbf{x}) & \text{for } y = -1 \end{cases} \tag{2}$$

and we know the maximum likelihood selects the hypothesis h, which maximizes the probability, which is equivalent to minimizing the quantity:

$$\frac{1}{N}\sum_{n=1}^{N} ln\left(\frac{1}{Pr\,(\,y\,|\,\mathbf{x}\,)}\right) \tag{3}$$

We can rewrite (2) in terms of indicator variables:

$$Pr\,\{\,y\,|\,\mathbf{x}\,\} = \mathbb{1}_{y\in\{+1\}}h(\mathbf{x}) + \mathbb{1}_{y\in\{-1\}}(1 - h(\mathbf{x})) \tag{4}$$

rewriting (3) in terms of (4) we obtain the following:

$$\frac{1}{N}\sum_{n=1}^{N} ln\left(\frac{1}{Pr\,(\,y\,|\,\mathbf{x}\,)}\right) = \frac{1}{N}\sum_{n=1}^{N} ln\left(\frac{1}{\mathbb{1}_{y\in\{+1\}}h(\mathbf{x}) + \mathbb{1}_{y\in\{-1\}}(1 - h(\mathbf{x}))}\right) \tag{5}$$

$$= \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}_{y\in\{+1\}} ln\left(\frac{1}{h(\mathbf{x})}\right) + \mathbb{1}_{y\in\{-1\}} ln\left(\frac{1}{1 - h(\mathbf{x})}\right) \tag{6}$$

which concludes the proof.

b

We want to prove that minimizing the in-sample error from **a** is equivalent to minimizing the following in-sample error:

$$E_{in}(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N} ln\left(1 + e^{-y_n\mathbf{w}^T\mathbf{x}_n}\right)$$

when $h(x) = \theta(\mathbf{w}^T\mathbf{x}) = \dfrac{e^{\mathbf{w}^T\mathbf{x}}}{1 + e^{\mathbf{w}^T\mathbf{x}}} = \dfrac{1}{1 + e^{-\mathbf{w}^T\mathbf{x}}}$

If we substitute this definition of $h(x)$ into (6) we get the following:

$$\mathbb{1}_{y\in\{+1\}} ln\left(\frac{1}{\theta(\mathbf{w}^T\mathbf{x})}\right) + \mathbb{1}_{y\in\{-1\}} ln\left(\frac{1}{1 - \theta(\mathbf{w}^T\mathbf{x})}\right)$$

From the second term we note that: $\theta(-x) = \dfrac{e^{-x}}{1 + e^{-x}} = \dfrac{1}{1 + e^{s}} = 1 - \theta(x)$ therefore $1 - \theta(x) = \theta(-x)$ and we can write:

$$\mathbb{1}_{y\in\{+1\}} ln\left(\frac{1}{\theta(\mathbf{w}^T\mathbf{x})}\right) + \mathbb{1}_{y\in\{-1\}} ln\left(\frac{1}{\theta(-\mathbf{w}^T\mathbf{x})}\right) = ln\left(\frac{1}{\theta(y_i\mathbf{w}^T\mathbf{x})}\right)$$

writing out the sigmoid function we get:

$$ln\left(\frac{1}{\theta(y_i\mathbf{w}^T\mathbf{x})}\right) = ln\left(\frac{1}{\frac{1}{1 + e^{-y_n\mathbf{w}^T\mathbf{x}_n}}}\right)$$

$$= ln\left(1 + e^{e^{-y_n\mathbf{w}^T\mathbf{x}_n}}\right)$$

thus we get the desired in-sample error:

$$E_{in}(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N} ln\left(1 + e^{-y_n \mathbf{w}^T \mathbf{x}_n}\right)$$

3.2 Logistic regression loss gradient

First we note that the in-sample error measure is defined as:

$$E_{in}(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N} ln\left(1 + e^{-y_n \mathbf{w}^T \mathbf{x}_n}\right)$$

We determine the gradient of the in-sample loss error measure:

$$\nabla E_{in}(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N} \frac{\partial}{\partial \mathbf{w}}\left[ln\left(1 + e^{-y_n \mathbf{w}^T \mathbf{x}_n}\right)\right]$$

If we let $f(x) = ln(x)$ and $g(x) = 1 + e^{-y_n \mathbf{w}^T \mathbf{x_n}}$, we apply the chain rule for gradients which means that we get:

$$\frac{\partial}{\partial \mathbf{w}}\left[ln\left(1 + e^{-y_n \mathbf{w}^T \mathbf{x}_n}\right)\right] = f'(g(\mathbf{w}))\nabla(\mathbf{w})$$

Applying above we have:

$$\begin{aligned} f'(g(\mathbf{w})) &= \frac{1}{1 + e^{-y_n \mathbf{w}^T \mathbf{x}_i}} \\ \nabla g(\mathbf{w}) &= \frac{\partial}{\partial \mathbf{w}}\left[1 + e^{-y_n \mathbf{w}^T \mathbf{x}_n}\right] = e^{-y_n \mathbf{w}^T \mathbf{x}_n} \times (-y_n \mathbf{x}_n) \end{aligned}$$

now we can compute $f'(g(\mathbf{w}))\nabla(\mathbf{w})$:

$$\begin{aligned} f'(g(\mathbf{w}))\nabla(\mathbf{w}) &= \frac{-y_n \mathbf{x}_n e^{-y_n \mathbf{w}^T \mathbf{x}_n}}{1 + e^{-y_n \mathbf{w}^T \mathbf{x}_n}} \\ &= \frac{-y_n \mathbf{x}_n e^{-y_n \mathbf{w}^T \mathbf{x}_n}/e^{-y_n \mathbf{w}^T \mathbf{x}_n}}{1 + e^{-y_n \mathbf{w}^T \mathbf{x}_n}/e^{-y_n \mathbf{w}^T \mathbf{x}_n}} \\ &= \frac{-y_n \mathbf{x}_n}{1 + e^{y_n \mathbf{w}^T \mathbf{x}_n}} \end{aligned}$$

Thus we can write:

$$\nabla E_{in}(\mathbf{w}) = -\frac{1}{N}\sum_{n=1}^{N} \frac{y_n \mathbf{x}_n}{1 + e^{y_n \mathbf{w}^T \mathbf{x}_n}}$$

It is clear that:

$$\frac{1}{N}\sum_{n=1}^{N} -y_n \mathbf{x}_n \theta(-y_n \mathbf{w}^T \mathbf{x}_n) = -\frac{1}{N}\sum_{n=1}^{N} \frac{y_n \mathbf{x}_n}{1 + e^{y_n \mathbf{w}^T \mathbf{x}_n}}$$

If we write out the sigmoid function:

$$\begin{aligned} \frac{1}{N}\sum_{n=1}^{N} -y_n \mathbf{x}_n \theta(-y_n \mathbf{w}^T \mathbf{x}_n) &= \frac{1}{N}\sum_{n=1}^{N} -y_n \mathbf{x}_n \frac{1}{1 + e^{y_n \mathbf{w}^T \mathbf{x}_n}} \\ &= \frac{1}{N}\sum_{n=1}^{N} \frac{-y_n \mathbf{x}_n}{1 + e^{y_n \mathbf{w}^T \mathbf{x}_n}} \\ &= -\frac{1}{N}\sum_{n=1}^{N} \frac{y_n \mathbf{x}_n}{1 + e^{y_n \mathbf{w}^T \mathbf{x}_n}} \end{aligned}$$

Thus completing the proof.

3.3 Logistic regression implementation

The implementation for logistic regression is found in the file `logRes.py`. The implementation works as followed:

1. First the data is handled. The features and target values are separated.The second class from the dataset is removed and the class 0 is changed to -1. The intercept is added for each feature: $\mathbf{x}_0 = 1$
2. Next the total number of iterations and the step size α for gradient decent is specified
3. For each iteration the gradient is calculated as:

   $$\mathbf{g}_t = -\frac{1}{N}\sum_{n=1}^{N}\frac{y_n\mathbf{w}^T}{1+e^{y_n\mathbf{w}^T(t)\mathbf{x}_n}}$$

   where the direction of the gradient is defined as $\mathbf{v}_t = -\mathbf{g}_t$
4. The weights are updated by: $\mathbf{w}(t+1) = \mathbf{w}(t) + \alpha\mathbf{v}_t$
5. After the last iteration the updated weights are returned.

Once the weights have been obtained the predictive class is calculated as:

$$h(x) = \theta(\mathbf{w}^T\mathbf{x})$$

where θ is the sigmoid function. The predictive class is the determined by:

$$h(x) = \begin{cases} +1 & \text{for } h(x) \geq 0.5 \\ -1 & \text{for } h(x) < 0.5 \end{cases}$$

The learning hypothesis is measured using a zero one less expressed as the empirical error:

$$\hat{L}(h) = \frac{1}{N}\sum_{n=1}^{N}\ell(h(x_n), y_n)$$

The below table shows the empirical error on the training and test data, that is the weights have been found using the training data and applied to the training and test data.

Test	Train
0.0168	0.076

The results from the above table have been generated by running gradient decent for 10000 iterations with a step size $\alpha = 0.1$ The weights found with the setting is

$$a = [\mathbf{w}_1, \mathbf{w}_2], b = [\mathbf{w}_0] :$$

$$a = [3.274, -14.386], b = [-13.015]$$