Advanced Topics in Machine Learning - Assignment 4

Christoffer Thrysøe - dfv107

October 3, 2017

1. SVM model selection

The RBF kernel is defined as:

$$K(x, x') = exp(-\gamma||x - x'||^2) \tag{1}$$

To see the transformed feature space, we can pick $\gamma = 1$ and write out the transformation:

$$\begin{aligned} K(x, x') &= exp(||x - x'||^2) && (2)\\ &= exp(-(x)^2) \cdot exp(2xx') \cdot exp(-(x')^2) && (3)\\ &= exp(-(x)^2) \cdot \sum_{k=0}^{\infty} \frac{2^k (x)^k (x')^k}{k!} \cdot exp(-(x')^2) && (4) \end{aligned}$$

Thus the non linear transform can be written as followed:

$$\Phi(x) = exp(-x^2) \cdot \left(1, \sqrt{\frac{2^1}{1!}}x, \sqrt{\frac{2^2}{2!}}x^2, \dots\right) \tag{5}$$

We have that the transformation Φ is an infinite dimensional transformation, therefore the data will most likely be separable in the transformed space. The parameter γ controls the width of the kernel . A small value of γ results in a wide kernel, while a large γ results in a narrow kernel. A narrow kernel will more likely lead to over-fitting as the geometric surface of the Φ-transformed feature space will be sharp, allowing the margin more freedom to separate the data, where a larger kernel will choose a larger margin, because the geometric surface is smooth. The parameter C is the regularization parameter and controls the penalty we add to the slack variables. A small value of C means that we allow more slack variables and thus can obtain a larger margin, whereas a large value for C will penalize slack variables, which will cause the SVM to classify as many data points correctly as possible, often resulting in a smaller margin and thus a worse generalization. When we perform hyper parameter selection, we can for each γ, arrange the values of C such that they are increasing. If the data, in the Φ-transformed space, is separable then there must exists some C' such that when we have $C > C'$, the SVM will be a hard margin SVM in the infinite dimensional feature space. In this case, any $C > C'$ will result in the same hyperplane, and therefore we can stop hyper parameter selection, because we will get the same hyperplane. It is clear to see from the SVM minimization problem shown in (6)

$$\text{minimize}_{\xi,w,b} \ \ \frac{1}{2}\langle w, w\rangle + C\sum_{i=1}^{l} \xi_i \tag{6}$$

If we are able to separate the data completely, we will get the following minimization problem:

$$\text{minimize}_{\xi,w,b} \quad \frac{1}{2}\langle w, w\rangle + C\sum_{i=1}^{l} 0 \tag{7}$$

the slack variable will be zero because all training examples are classified correctly and clearly increasing C will not result in a different hyperplane. Thus we know that when the error on the training data is zero, increasing C will not change the margin and we can therefore move on to the next value of γ.
I believe that the above can also be used for inseparable data in the Φ-transformed space. At some point the penalty term, C will make the hyperplane separate as many data points as possible, but because the data is inseparable in the Φ-transformed space, the hyperplane will be fixed when increasing C. If we have that the weights of the hyperplane don't change, when increasing C, we will know that we can stop, as we will get the same hyperplane regardless of the increase in C. This stopping criteria will also apply to when the data is separable.

2. Least Squares SVMs

For the regular L_2-norm, soft margin SVM we have the following minimization problem:

$$\text{minimize}_{\xi,w,b} \quad \frac{1}{2}\langle w, w\rangle + \frac{C}{2}\sum_{i=1}^{l} \xi_i^2 \tag{8}$$

$$\text{subject to} \quad y_i(\langle w, \Phi(x_i)\rangle + b) \geq 1 - \xi_i \tag{9}$$

Instead of the hinge loss, the least squares SVM uses the following error function:

$$L(y_i, f(x_i)) = (y_i - f(x_i))^2 = e_i^2 \tag{10}$$

The least squares SVM can then be formulated in the following way:

$$\text{minimize}_{w,b,e} \quad \frac{1}{2}\langle w, w\rangle + \frac{\gamma}{2}\sum_{i=1}^{l} e_i^2 \tag{11}$$

1

As presented in the lecture notes, we can reformulate the problem as a regularized risk minimization:

$$\text{minimize}_{f\in H_k^b} \quad \frac{1}{l}\sum_{i=1}^{l} L(y_i, f(x_i)) + \upsilon_l ||f||_k^2 \tag{12}$$

where $\upsilon_l = (l\gamma)^{-1}$. If we use the least squares error function, we get the following representation:

$$\text{minimize}_{f\in H_k^b} \quad \frac{1}{l}\sum_{i=1}^{l} (y_i - f(x_i))^2 + \upsilon_l ||f||_k^2 \tag{13}$$

next we can use the Representer theorem to get a representation of $f(x)$ which has the form:

$$f(x) = \sum_{i=1}^{l} \alpha_i k(x_i, x) \tag{14}$$

which will give us the following:

$$\min_{\alpha} \quad \frac{1}{l}\sum_{i=1}^{l} (y_i - \sum_{j=1}^{l} \alpha_j k(x_i, x_j))^2 + \upsilon_l ||\sum_{j=1}^{l} \alpha_j k(x_j, \cdot)||_k^2 \tag{15}$$

which is a problem we can minimize, by taking the partial derivative with respect to α, setting it to zero and finding the solution to α.

2

The constraint on the least squares SVM minimization problem from (11) is the following equality constraint:

$$y_i[w^T\Phi(x_i) + b] = 1 - e_i \tag{16}$$

3. Regularization by relative entropy and the Gibbs distribution

Given the following minimization problem:

$$\min_{\rho_1,\ldots,\rho_m} \quad \alpha\sum_{h=1}^{m}\rho_h L_h + \sum_{h=1}^{m}\rho_h \ln\frac{\rho_h}{\pi_h} \tag{17}$$

$$s.t. \quad \sum_{h=1}^{m}\rho_h = 1 \tag{18}$$

$$\forall h : \rho_h \geq 0 \tag{19}$$

we wish to show that the solution has the following form

$$\rho_h = \frac{\pi_h e^{-\alpha L_h}}{\sum_{h'=1}^{m}\pi_{h'}e^{-\alpha L'_h}} \tag{20}$$

To show this, I will follow the given guidelines and drop the last constraint and prove it later. First we write up the Lagrange function, where the equality constraint is changed to be equal to zero:

$$L(\rho,\lambda) = \alpha\sum_{h=1}^{m}\rho_h L_h + \sum_{h=1}^{m}\rho_h \ln\frac{\rho_h}{\pi_h} + \lambda(\sum_{h=1}^{m}\rho_h - 1) \tag{21}$$

$$= \alpha\sum_{h=1}^{m}\rho_h L_h + \sum_{h=1}^{m}\rho_h \ln\rho_h - \sum_{h=1}^{m}\rho_h \ln\pi_h + \lambda(\sum_{h=1}^{m}\rho_h - 1) \tag{22}$$

Now we take the derivative of (22) with respect to each ρ_h and set them to zero. For a single ρ_h this is given by:

$$\nabla L_{\rho_h}(\rho_h,\lambda) = \alpha L_h + \ln\rho_h + 1 - \ln\pi_h + \lambda = 0 \tag{23}$$

From this we get:

$$\ln\rho_h = \ln\pi_h - \alpha L_h - 1 - \lambda \Rightarrow \tag{24}$$

$$\rho_h = \pi_h e^{-\alpha L_h - 1 - \lambda} = \pi_h e^{-\alpha L_h}e^{-1-\lambda} \tag{25}$$

π is a (prior) distribution, therefore we have $\pi_h \in [0,1]$, looking at (25) we can easily verify that $\rho_h \geq 0$ for all h, because e^{-x} will be greater than 0 for all x and the term can therefore never be negative.
Since we have given that the distribution ρ must sum to 1, we can use (25) to get the following:

$$\sum_{h=1}^{m}\rho_h = e^{-1-\lambda}\sum_{h=1}^{m}\pi_h e^{-\alpha L_h} = 1 \tag{26}$$

rewriting (26), we get:

$$e^{-1-\lambda} = \frac{1}{\sum_{h=1}^{m}\pi_h e^{-\alpha L_h}} \tag{27}$$

plugging this value back into (25) we have the following value of ρ_h

$$\rho_h = \pi_h e^{-\alpha L_h}\frac{1}{\sum_{h=1}^{m}\pi_h e^{-\alpha L_h}} \tag{28}$$

$$= \frac{\pi_h e^{-\alpha L_h}}{\sum_{h=1}^{m}\pi_h e^{-\alpha L_h}} \tag{29}$$

Which is equal to (20) and therefore what we wanted to prove.

4. Follow The Leader (FTL) algorithm for i.i.d. full information games

For the "Follow The Leader" approach, at each time t we play the arm which was the most successful up to round t. That is, at each round we play the arm:

$$A_t = \max_a \hat{\mu}_{t-1}(a) \tag{30}$$

Because the game is played in the full information setting, we will know for each time t, which arm satisfies (30). We wish to bound the pseudo regret of the "Follow the Leader" strategy. First we note that the number of times an action a was played can be written as followed:

$$N_{T(a)} = \sum_{t=1}^{T} 1\{A_t = a\} \tag{31}$$

where

$$\mathbb{E}[1\{A_t = a\}] \leq P\{\hat{\mu}_{t-1}(a) \geq \max_{a'} \hat{\mu}_{t-1}(a')\} \tag{32}$$

$$\leq P\{\hat{\mu}_{t-1}(a) \geq \hat{\mu}_{t-1}(a^*)\} \tag{33}$$

From the lecture notes, we have that the pseudo regret $\bar{R}_T$ is defined as:

$$\bar{R}_T = \sum_a \Delta(a)\mathbb{E}[N_T(a)] \tag{34}$$

Thus if we can bound (33), we can get a bound on the pseudo regret. To bound (33) I will use the same approach as in the lecture notes, but for full information, instead of bandit setting:

$$P\{\hat{\mu}_{t-1}(a) \geq \hat{\mu}_{t-1}(a^*)\} \tag{35}$$

$$\leq P\{\mu_{t-1}(a) \geq \hat{\mu}(a) + \frac{1}{2}\Delta\} + P\{\hat{\mu}_{t-1}(a^*) \leq \mu^* - \frac{1}{2}\Delta\} \tag{36}$$

$$\leq 2e^{-2t-1(\frac{1}{2}\Delta)^2} = 2e^{-t-1\Delta^2/2} \tag{37}$$

where the last inequality is from Hoeffdings inequality. Thus we have the bound:

$$P\{\hat{\mu}_{t-1}(a) \geq \hat{\mu}_{t-1}(a^*)\} \leq 2e^{-\frac{\Delta^2}{2}(t-1)} \tag{38}$$

Now we can define $\mathbb{E}[N_T(a)]$:

$$\mathbb{E}[N_T(a)] = \sum_{t=1}^{T} \mathbb{E}[1\{A_t = a\}] \tag{39}$$

$$\leq \sum_{t=1}^{T} P\{\hat{\mu}_{t-1}(a) \geq \hat{\mu}_{t-1}(a^*)\} \tag{40}$$

$$\leq \sum_{t=1}^{T} 2e^{-\frac{\Delta^2}{2}(t-1)} \tag{41}$$

$$= 2\sum_{t=1}^{T} e^{-\frac{\Delta^2}{2}(t-1)} = 2\sum_{t=0}^{T} e^{-\frac{\Delta^2}{2}t} \tag{42}$$

when taking $t \to \infty$ over (42), we sum up a geometric series, which is defined as followed:

$$\sum_{t=0}^{\infty} r^t = \frac{1}{1-r} \tag{43}$$

for $r < 1$. In our case we have $r = e^{-\frac{\Delta^2}{2}}$ which is always less than 1, when $\Delta(a) > 0$. Thus we get the following:

$$\mathbb{E}[N_t(a)] \leq 2 \cdot \frac{1}{1 - e^{-\frac{\Delta^2}{2}}} \tag{44}$$

combining this with the definition on the pseudo regret from (34), we get the following bound:

$$\bar{R}_T \leq \sum_{a:\Delta(a)>0} \frac{2}{1 - e^{-\frac{\Delta(a)^2}{2}}} \Delta(a) \tag{45}$$

which is the desired bound.