



Foundations of Machine Learning (CS 725)

FALL 2024

More Practice Questions + Course Conclusion

Question 1

Logistic regression extended to K classes is called softmax regression. In this case, $y \in \{0, 1, 2, \dots, K - 1\}$ and the posterior probability that an example \mathbf{x} belongs to class k is defined as:

$$P(y = k|\mathbf{x}) = \frac{\exp(\mathbf{w}_k^T \mathbf{x})}{\sum_{j=1}^K \exp(\mathbf{w}_j^T \mathbf{x})}$$

where \mathbf{w}_k is a weight vector corresponding to the k^{th} class. Given N training examples $\{\mathbf{x}_i, y_i\}_{i=1}^N$, fill in the blanks to complete the expression for the negative log likelihood of the data:

$$-\log \prod_{i=1}^N P(y_i|\mathbf{x}_i) = -\sum_{i=1}^N \sum_{k=1}^K (\delta(y_i = k) \cdot \underline{\hspace{2cm}}) + \sum_{i=1}^N \log \underline{\hspace{2cm}}$$

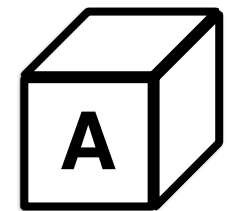
Solution:

$$\text{where } \delta(y_i = k) = \begin{cases} 1 & \text{if } y_i = k \\ 0 & \text{if } y_i \neq k \end{cases}$$

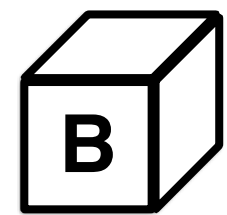
$$\begin{aligned} -\log \prod_{i=1}^N P(y_i|\mathbf{x}_i) &= -\log \prod_{i=1}^N \prod_{k=1}^K \left(\frac{\exp(\mathbf{w}_k^T \mathbf{x}_i)}{\sum_{j=1}^K \exp(\mathbf{w}_j^T \mathbf{x}_i)} \right)^{\delta(y_i=k)} \\ &= -\sum_{i=1}^N \sum_{k=1}^K \delta(y_i = k) \mathbf{w}_k^T \mathbf{x}_i + \sum_{i=1}^N \log \left(\sum_{j=1}^K \exp(\mathbf{w}_j^T \mathbf{x}_i) \right) \end{aligned}$$

Question 2

Consider the space of all possible subsets of a given fixed set \mathcal{D} . Prove or disprove the following functions are valid kernels. Let A_1 and A_2 be subsets of \mathcal{D} .

 $K(A_1, A_2) = |A_1 \cap A_2|$

Solution: Construct a feature vector of size of \mathcal{D} with 0 or 1 for an element of a set being absent or present respectively in a subset. Taking the product of two feature vectors would give the number of common elements because only the 1's corresponding to elements present in both subsets refer to the intersection of both subsets.

 $K(A_1, A_2) = 2^{|A_1 \cap A_2|}$

Solution: One way of showing this is to show how if $|A_1 \cap A_2|$ is a kernel, then $2^{|A_1 \cap A_2|}$ is also a kernel. We can also show implicit feature vectors that make this a valid kernel. Consider a feature vector Φ of size $2^{|\mathcal{D}|}$. Then $\Phi_U(A_1) = \begin{cases} 1 & \text{if } U \subseteq A_1 \\ 0 & \text{otherwise} \end{cases}$

Question 3

The linear combination of two kernels $K_1(x_1, x_2)$ and $K_2(x_1, x_2)$, with implicit feature vectors $\Phi_1(x)$ and $\Phi_2(x)$, is $K(x_1, x_2) = aK_1(x_1, x_2) + bK_2(x_1, x_2)$, $a \geq 0, b \geq 0$ which is also a Mercer kernel. What would be the implicit feature vector $\Phi(x)$ for K in terms of $\Phi_1(x)$ and $\Phi_2(x)$?

Solution: $\Phi(x) = [\sqrt{a}\Phi_1(x), \sqrt{b}\Phi_2(x)]$

Question 4

Consider the Gaussian kernel $K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{1}{2}||\mathbf{x} - \mathbf{x}'||^2\right)$. Let ϕ be the feature function induced by K . What does $||\phi(\mathbf{x}) - \phi(\mathbf{x}')||^2 + ||\phi(\mathbf{x}) + \phi(\mathbf{x}')||^2$ evaluate to?

Solution: Answer is 4.

$$||\phi(\mathbf{x}) - \phi(\mathbf{x}')||^2 = (\phi(\mathbf{x}) - \phi(\mathbf{x}'))^T (\phi(\mathbf{x}) - \phi(\mathbf{x}')) = ||\phi(\mathbf{x})||^2 + ||\phi(\mathbf{x}')||^2 - 2\phi(\mathbf{x})^T \phi(\mathbf{x}')$$

Similarly, $||\phi(\mathbf{x}) + \phi(\mathbf{x}')||^2 = ||\phi(\mathbf{x})||^2 + ||\phi(\mathbf{x}')||^2 + 2\phi(\mathbf{x})^T \phi(\mathbf{x}')$. $||\phi(\mathbf{x})||^2 = K(\mathbf{x}, \mathbf{x}) = 1$ and $||\phi(\mathbf{x}')||^2 = K(\mathbf{x}', \mathbf{x}') = 1$. Hence, the answer adds up to 4.

Question 5

Recall the perceptron weight update rule on a given example (x, y) , $x \in \mathbb{R}^d$, $y \in \{+1, -1\}$, when a misclassification occurs:

$$w_{t+1} \leftarrow w_t + \eta y x.$$

Suppose we present each example (x, y) repeatedly to the perceptron algorithm until it correctly classifies it. Give an upper bound on the number of times an example will be repeated, given that when the first example was presented the weight was w_i . (The upper bound can be in terms of η , w_i and (x, y) .)

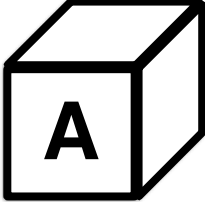
Solution: Find the largest k such that $yw_{i+k-1}^T x = yw_i^T x + \eta(k-1)y^2\|x\|^2 < 0$

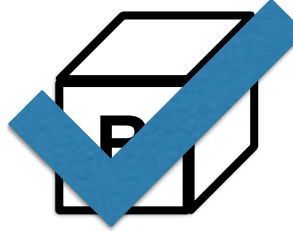
$$\Rightarrow k < \frac{-yw_i^T x}{\eta\|x\|^2} + 1$$

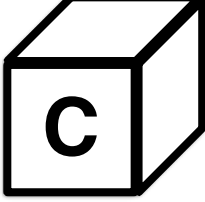
Since k is an integer, we have $k < \left\lceil \frac{-yw_i^T x}{\eta\|x\|^2} \right\rceil$

Question 6

Consider a DT with n leaves, for a binary classification task (with labels “+” and “−”). Suppose a dataset is such that, for $i = 1$ to n , a p_i fraction of them fall into the i^{th} leaf. Also suppose that of the samples which fall into the i^{th} leaf, a fraction q_i have the positive label. Suppose you assign labels to the leaves of the DT in order to minimize the classification error on this dataset. Choose the right expression(s) for the resulting error rate from below. Justify your answer.

 $\sum_{i=1}^n p_i q_i + (1 - p_i) \left| \frac{1}{2} - q_i \right|$

 $\sum_{i=1}^n p_i \left(\frac{1}{2} - \left| \frac{1}{2} - q_i \right| \right)$

 $\sum_{i=1}^n p_i \left(\frac{q_i}{2} + \left| \frac{1}{2} - q_i \right| \right)$

Solution: To minimize classification error, assign the label “+” to the i^{th} leaf iff $q_i \geq 0.5$. The resulting error rate among these samples falling into the i^{th} leaf is:

$$e_i = \begin{cases} 1 - q_i & \text{if } q_i \geq 0.5 \\ q_i & \text{if } q_i < 0.5 \end{cases}$$

But if $q_i \geq 0.5$, $1 - q_i = 0.5 - (q_i - 0.5) = 0.5 - |0.5 - q_i|$

and if $q_i < 0.5$, $q_i = 0.5 - (0.5 - q_i) = 0.5 - |0.5 - q_i|$

So, error rate = $\sum_i p_i e_i = \sum_i p_i \left(\frac{1}{2} - \left| \frac{1}{2} - q_i \right| \right)$

Course Remnants



- Final exam: Nov 23, 2024
(5:30 pm to 8:30 pm, Venues: LA001, LC001)
- Showing final exam answer sheets: Nov 25-26, 2024
- Project Evaluations: Nov 25-26, 2023

Thanks!

By the way, ChatGPT misspelled your name.