


## Homework 5

### (Probabilistic Modeling and Decision Theory)

For questions, please refer to Moodle.  
Released on **10 May 2022**

#### GENERAL INSTRUCTIONS


- Submission of solutions is not mandatory but solving the exercises are highly recommended. The master solution will be released next week.
- Part of the exercises are available on Moodle as a quiz. These problems are marked with .

### Exercise 1: Multiclass logistic regression


The posterior probabilities for multiclass ( $K$  classes) logistic regression can be given as a softmax transformation of hyperplanes, such that:

$$f_k(y, x, a_1, \dots, a_K) = P(y = k \mid X = \mathbf{x}, a_1, \dots, a_K) = \frac{\exp(\mathbf{a}_k^T \mathbf{x})}{\sum_j \exp(\mathbf{a}_j^T \mathbf{x})}$$


Here,  $x$  and  $a_i$  are vectors of size  $d$ . Let's also define  $s_i = \mathbf{a}_i^T x$  for all  $1 \leq i \leq K$ . Now, let's examine the derivative  $D_{ki} = \frac{\partial f_k}{\partial s_i}(y, x, a_1, \dots, a_K)$ .

(a)  What shape does  $D_{ki}$  have?

1. Scalar
2. Vector of size  $d$
3. Vector of size  $k$
4. Matrix of size  $d \times K$  (or  $K \times d$ )
5. Depends on  $k$  and  $i$


(b)  What is a value of  $D_{ki}$ ?

1.  $f_k(\mathbb{1}\{k = i\}) - f_i$
2.  $-f_k(\mathbb{1}\{k = i\}) - f_i$
3.  $f_k - f_i$
4.  $f_k - f_k f_i$

(c)  In practice, we want to have the derivative by  $a_i$  instead of  $s_i$  (why?). What will this derivative look like?

$$1. D_{ki} x_i \quad 2. D_{ki} x_k \quad 3. x_i D_{ki} \quad 4. x_k D_{ki} \quad 5. \sum_j x_j D_{kj}$$

Note that it was enough to compute exponents for each class once to derive  $f_i$ . Once we have  $f_i$ , derivatives reuse it. However, still, some problems may appear. Let's find them and solve:

(d)  What will happen if some  $s_i$  is large (e.g.,  $> 1000$ )?

1. Nothing, because  $f_i$  is still bounded from 0 to 1, so we are safe here.
2. Nothing, because I ran `scipy.special.softmax`, and everything works just fine.
3. Nothing, because Python uses big integers, and  $\exp(s_i)$  will take some time to compute, but it will work.

#### 4. Overflow when counting $\exp(s_i)$

Imagine that you live in the real world and intermediate results in computation matter. Also, imagine that you don't want to use big integers for many reasons (performance is one of them). But scipy indeed can compute softmax even with large  $s_i$ . To achieve this, let's apply a trick.

- (e) [✓] In the lecture, it was mentioned that we might set one of the vectors as a null vector to guarantee uniqueness. Instead of changing weights, let's set one  $s_{i^*}$  to zero (only for calculating  $f_k$ : for derivative computation, there is no need for this). What should be the initial (before setting to 0) value of  $s_{i^*}$  and how to change other  $s_j$  to have all the  $f_k$  unchanged?
1.  $s_{i^*} = \min_{i \in K} s_i$ . Divide other weights by  $s_{i^*}$ .
  2.  $s_{i^*} = \min_{i \in K} |s_i|$ . Subtract  $s_{i^*}$  from other weights.
  3.  $s_{i^*} = \min_{i \in K} |s_i|$ . Divide other weights by  $s_{i^*}$ .
  4.  $s_{i^*} = \max_{i \in K} s_i$ . Subtract  $s_{i^*}$  from other weights.
  5.  $s_{i^*} = \max_{i \in K} s_i$ . Divide  $s_{i^*}$  from other weights.

## Exercise 2: Decision Theory

A *binary option* is a simple derivative, which gives you some fixed profit if you predicted something right (e.g., the stock of company ABC at noon on the 24th of April will be higher than 20\$) or gives you nothing otherwise. An investor considers buying a binary option for 1000 CHF that can make a profit of 600 CHF (payoff is 1600 including his investment). Alternatively, he can also invest the money securely, in which case he has a guaranteed interest of 200 CHF over the same time. The investor uses logistic regression to predict the probability of a binary option making a profit.  $x \in \mathbb{R}^d$  encodes the information about the binary option.

- (a) What is the estimated conditional probability of profit from the binary option? Define the action set  $\mathcal{A}$  and cost function  $\mathcal{C} : \mathcal{Y} \times \mathcal{A} \rightarrow \mathbb{R}$ .
- (b) [✓] What is the expected cost if the investor decides to buy a binary option?
- (c) [✓] Assuming the model is correct, what should be a decision rule for buying (or not) a binary option? Obviously, the decision function should look like:

$$f(x) = \begin{cases} \text{secure investment} & p(x) \leq p^* \\ \text{buy option} & p(x) > p^* \end{cases}$$

for some threshold  $p^*$ . In other words, the investor will buy a binary option i.f.f., probability that option will make him profit is above  $p^*$ . What is  $p^*$ ?

- (d) What happens if the model is not correct? Let's consider a blackbox model that only outputs a binary outcome but is only correct with probability  $p$ . How will the decision function change?

## Exercise 3: Naive Bayes Estimate

In this task we will use the Naive Bayes model for binary classification. Let  $\mathcal{Y} = \{0, 1\}$  be the set of labels and  $\mathcal{X} = \mathbb{N}^d$  a  $d$ -dimensional features space ( $\mathbb{N} = \{0, 1, 2, \dots\}$ ). You are given a training set  $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$  of  $n$  labeled examples  $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ , where all features are geometric distributed with parameters  $\hat{p}_j$  for  $j \in \{1, 2, \dots, d\}$ .

- (a) Let  $\{z_i\}_{i=1}^m$  be  $m$  i.i.d. observations of a  $p$ -geometric distributed random variable. Find the maximum likelihood estimation of  $p$  for this model. (Hint: For a  $p$ -geometric distributed random variable  $Z$ , we have that  $P[Z = K] = p(1 - p)^{K-1}$ , for a  $K \in \mathbb{N}$ .)





- (b) We now want to train a geometric naive bayes solution using the maximum likelihood estimates. Define appropriate parameters  $\hat{p}_{j,0}, \hat{p}_{j,1} \in [0, 1]$  and  $p_0, p_1 \in [0, 1]$ , and write down the joint distribution  $P(X, Y)$  of the resulting naive bayes model. (Hint:  $p_0 + p_1 = 1$ )
- (c) Next, we want to use our model of (b) to minimize the misclassification probability of a new sample  $x \in \mathcal{X}$ , i.e.  $y_{pred} = \arg \max_{y \in \mathcal{Y}} P[y | X = x]$ . Show that the predicted label  $y_{pred}$  for  $x$  is determined by a hyperplane, i.e., that  $y_{pred} = [a^\top x \geq b]$  for some  $a \in \mathbb{R}^d, b \in \mathbb{R}$ .
- (d) The naive Bayes model assumes that all covariates are independent. This, however, may be a very strong and false assumption. For example, a dataset that contains the age of a person and their year of birth has a perfect correlation between these attributes. For visualization, we look at the following example. Assume that the features 2 to  $d$  of our data model are all equal ( $x_{i,2} = x_{i,j}, \forall j \in \{3, \dots, d\}$ ), i.e. perfectly correlated, but the first feature is independent of the rest. What should the decision plane be according to the rule  $y_{pred} = \arg \max_{y \in \mathcal{Y}} P[y | X = x]$ ?

## Exercise 4: True or False: bias-variance trade-off

Assume we have a dataset of  $n$  i.i.d. samples  $\{x_i, y_i\}_{i=1}^n$  and we consider least squares regression, i.e.  $\hat{h} = \arg \min_{h \in \mathcal{H}} \sum_{(x,y) \in D} (y - h(x))^2$ . In class we have learned about the bias-variance trade-off, where we saw that

$$\text{prediction error} = \text{bias}^2 + \text{variance} + \text{noise} \quad (1)$$

For the following statements indicate true or false.

- |   |  |  |
|---|--|--|
| (a)  | It is impossible to reduce the prediction error if the bias increases.   | <input type="checkbox"/> True <input type="checkbox"/> False |
| (b)  | For an increasing amount of data, the variance decreases, but the bias increases.  | <input type="checkbox"/> True <input type="checkbox"/> False |
| (c)  | If we change to a strictly larger hypothesis class $\mathcal{H}_{old} \subset \mathcal{H}_{new}$ , then the bias must be smaller or equal to the bias considering the old hypothesis class $\mathcal{H}_{old}$ . | <input type="checkbox"/> True <input type="checkbox"/> False |
| (d)  | If $n \rightarrow \infty$ and the finite hypothesis class includes the optimal model $h^*$ , then the prediction error equals a term only dependent on the noise and the optimal hypothesis.                     | <input type="checkbox"/> True <input type="checkbox"/> False |

## Exercise 5: True or False: Distribution shifts

An ambitious PhD student wants to train a classifier, which can predict, given a tissue-patch, if the patient has cancer or not. In Figure 1, we see a subset of the data that the PhD student gathered (hospitals 1 to 3). We see an example of a tumorous tissue and of a normal tissue for each hospital she visited. Unfortunately for her, it was due to privacy regulations, impossible to gather data from hospital 4. In fact, it is infeasible to obtain data from all hospitals, but as our PhD student is highly ambitious: she would like her model to work independent of the hospital where the tissue sample was taken. For the following statements indicate true or false.

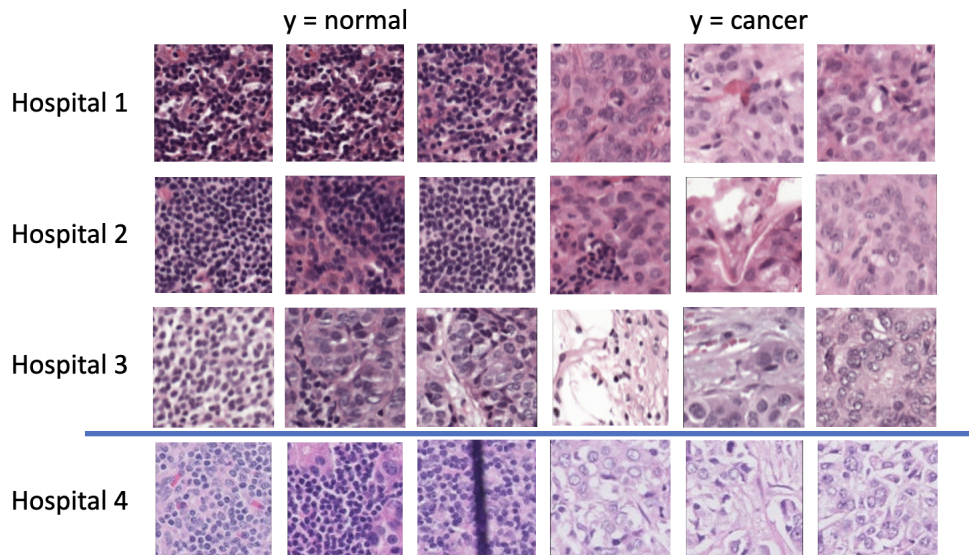


Figure 1: Images from the Camelyon17-wilds dataset [1], which comprises of tissue patches from different hospitals.

- In Figure 1, we see examples of tissue patches from hospital 4. Like the tissue-patches of hospitals 1,2 and 3, the tissue-patches from hospital 4 contain all necessary information to classify if the respective patient has cancer or not. Hence, the PhD student can use data from hospitals 1 to 3 to train a model that is also accurate on tissues-patches from hospital 4.
- (a) ☒ True ☐ False
- The assumption that the PhD student has i.i.d. distributed data of her target distribution to detect cancer given tissue patches of any hospital is close to the truth. Hence, her model will have equal classification error on new data from hospital 1 to 3 as on new data from hospital 4.
- (b) ☒ True ☐ False
- The assumption that she has i.i.d. distributed data to detect tissue samples from hospitals 1 to 3 is close to the truth. Therefore, her model will likely have a lower classification error on tissue-patches coming from hospitals 1 to 3, than from hospital 4.
- (c) ☒ True ☐ False
- The professor of the PhD student advises her to only work with i.i.d. data from her target distribution. Hence, she needs to partition her data according to hospital and train 3 different classifiers, one for each hospital.
- (d) ☒ True ☐ False

## References

- [1] Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE Transactions on Medical Imaging*, 2018.