Exercises
Introduction to Machine Learning
SS 2019

Series 6, May 6th, 2019
(Decision Theory, Logistic Regression)

Institute for Machine Learning
Dept. of Computer Science, ETH Zürich
Prof. Dr. Andreas Krause
Web: `https://las.inf.ethz.ch/teaching/introml-s19`
Email questions to:
natalie.davidson@inf.ethz.ch, marc.roeschlin@inf.ethz.ch

Solutions will be published on Monday, May 13th 2019.

Problem 1 (Decision Theory):

In this task, you would like to classify whether an X-ray result is cancerous or normal. The cost for a correct classification is 0 and the cost for predicting that the X-ray is normal when the true label is cancer is 1000, and the cost for predicting the X-ray is cancerous when the true label is normal is 1.

(i) Write out the cost function, estimated conditional distribution, and the action set. Justify why we would introduce an asymmetric cost.

(j) Write the action that will minimize the expected cost.

Problem 2 (Poisson Naive Bayes):

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 $D = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$ of n labeled examples $(\mathbf{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y}$.

(i) Is the Naive Bayes model a generative or a discriminative model? Justify your answer.

(ii) Let λ be a positive scalar, and assume that $z_1, \dots, z_m \in \mathbb{N}$ are m iid observations of a λ-Poisson distributed random variable. Find the maximum likelihood estimator for λ in this model. *(Hint: A λ-Poisson distributed random variable Z takes values $k \in \mathbb{N}$ with probability $P(Z = k) = e^{-\lambda} \frac{\lambda^k}{k!}$.)*

(iii) Let's train a Poisson Naive Bayes classifier using maximum likelihood estimation. Define appropriate parameters $p_0, p_1 \in [0, 1]$, and vectors $\lambda_0, \lambda_1 \in \mathbb{R}^d$, and write down the joint distribution $P(X, Y)$ of the resulting model. *(Note that the following should be satisfied for the parameters: $p_0 + p_1 = 1$, and λ_0, λ_1 are vectors with non-negative components.)*

(iv) Now, we want to use our trained model from (iii) to minimize the misclassification probability of a new observation $\mathbf{x} \in \mathcal{X}$, i.e., $y_{\text{pred}} = \operatorname{argmax}_{y \in \mathcal{Y}} P(y|X = \mathbf{x})$. Show that the predicted label y_{pred} for $\mathbf{x}$ is determined by a hyperplane, i.e., that $y_{\text{pred}} = \left[\mathbf{a}^\top \mathbf{x} \geq b\right]$ for some $\mathbf{a} \in \mathbb{R}^d, b \in \mathbb{R}$.

(v) Instead of simply predicting the most likely label, one can define a cost function $c : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, such that $c(y_{\text{pred}}, y_{\text{true}})$ is the cost of predicting y_{pred} given that the true label is y_{true}. Define the Bayes optimal decision rule for a cost function $c(\cdot, \cdot)$, with respect to a distribution $P(X, Y)$.

(vi) Write down a cost function such that the corresponding decision rule that you have defined in (v) for this cost coincides with a decision rule that minimizes the misclassification probability, i.e., $y_{\text{pred}} = \operatorname{argmax}_{y \in \mathcal{Y}} P(y|X = \mathbf{x})$.

Problem 3 (Multiclass logistic regression):

The posterior probabilities for mulitclass logistic regression can be given as a softmax transformation of hyperplanes, such that:

$$P(y = k|X = \mathbf{x}) = \frac{\exp(\mathbf{a}_k^\top \mathbf{x})}{\sum_j \exp(\mathbf{a}_j^\top \mathbf{x})}$$

If we consider the use of maximum likelihood to determine the parameters $\mathbf{a}_k$, we can take the negative logaritm of the likelihood function to obtain the *cross-entropy* error function for multiclass logistic regression:

$$E(\mathbf{a}_1, \ldots, \mathbf{a}_K) = -\ln \left(\prod_{n=1}^{N} \prod_{k=1}^{K} P(y = k|X = \mathbf{x}_n)^{t_{nk}} \right) = -\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \ln P(y = k|X = \mathbf{x}_n)$$

where $t_{nk} = 1_{[\text{labelOf}(\mathbf{x}_n) = k]}$.

Show that the gradient of the error function can be stated as given below *(refer to Bishop p. 209)*:

$$\nabla_{\mathbf{a}_k} E(\mathbf{a}_1, \ldots, \mathbf{a}_K) = \sum_{n=1}^{N} \left[P(y = k|X = \mathbf{x}_n) - t_{nk} \right] \mathbf{x}_n$$