Exercises
Introduction to Machine Learning
FS 2020

Institute for Machine Learning
Dept. of Computer Science, ETH Zürich
Prof. Dr. Andreas Krause
Web: `https://las.inf.ethz.ch/teaching/introml-s20`
For questions, please refer to Piazza.

Series 6, May 16th, 2020 (Decision Theory, Logistic Regression)

Problem 1 (Decision Theory):

In this task, you would like to classify whether an X-ray result is cancerous or normal, using a logistic model. 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. Answer the questions based on this task. The notation used in the questions is as follows: $\mathbf{x}$: X-ray features of a specific data point y : The label of a specific data point

$$y = \begin{cases} 0 & \text{if the sample is benign} \\ 1 & \text{ if the sample is cancerous} \end{cases}$$

X, Y : random variables denoting the X-ray features and the label, respectively a: Predicted label/action given X-ray features, x $\sigma(x) = \frac{1}{1+e^{-x}}$ $\mathbf{w}$: weight vector parameterising the logistic regression model $p = P(Y = 1|X = \mathbf{x})$

1. Pick the *action set* for the task.

   (a) A = Cancerous = 1, Benign = -1, Unknown = 0

   (b) A = Cancerous = 1, Benign = 0

   (c) A = Cancerous, given that the true label is cancerous = 0; Cancerous, given that the true label is benign = 1; Benign, given that the true label is cancerous = 1000; Benign, given that the true label is benign = 0

   (d) A = Cancerous = -2, Benign = 2, Not cancerous = 1, Not benign = -1

   Solution:
   The correct answer is (b).
   The action set for the prediction task is A = Cancerous = 1, Benign = 0.

2. Estimate the conditiional distribution of y, which determines the action.

   (a) $Bernoulli(y : \sigma(\mathbf{w}^T\mathbf{x}))$

   (b) $Bernoulli(a : \sigma(\mathbf{w}^T y))$

   (c) $Bernoulli(a : \sigma(\mathbf{w}^T\mathbf{w}))$

   (d) $Bernoulli(y : \sigma(\mathbf{x}^T\mathbf{x}))$

   Solution:
   The correct answer is (a).
   The conditional distribution, $P(y|\mathbf{x};\mathbf{w}) = Bernoulli(y : \sigma(\mathbf{w}^T\mathbf{x}))$.

3. Pick the correct cost function.

(a)

$$f(\mathbf{x}) = \begin{cases} 0 & \text{If the label is correct} \\ 1 & \text{If classified benign sample as cancerous} \\ 1000 & \text{If classified cancerous sample as benign} \end{cases}$$

(b)

$$f(\mathbf{x}) = \begin{cases} 0 & \text{If the label is correct} \\ 1 & \text{If classified cancerous sample as benign} \\ 1000 & \text{If classified benign sample as cancerous} \end{cases}$$

(c)

$$f(\mathbf{x}) = \begin{cases} 0 & \text{If the label is correct} \\ 1 & \text{If classified benign sample as cancerous} \\ 1 & \text{If classified cancerous sample as benign} \end{cases}$$

(d)

$$f(\mathbf{x}) = \begin{cases} 0 & \text{If the label is correct} \\ 1000 & \text{If classified benign sample as cancerous} \\ 1000 & \text{If classified cancerous sample as benign} \end{cases}$$

Solution:
The correct answer is (a).
This follows from the question description.

4. Pick the action that will minimize the expected cost. Try to prove the same.

   (a) Label the sample cancerous when $P(Y = 1|\mathbf{x}) > 1/1001$
   (b) Label the sample cancerous when $P(Y = 1|\mathbf{x}) > 1/1000$
   (c) Label the sample cancerous when $P(Y = 0|\mathbf{x}) > 1/1001$
   (d) Label the sample cancerous when $P(Y = 0|\mathbf{x}) > 1/1000$

Solution:
The correct answer is (a).
Let $C(Y, a)$ be the cost when the true label is Y and the action is a.

$$C(Y, a) = \begin{cases} 0 & \text{If Y=a} \\ 1 & \text{If Y=0, a=+1} \\ 1000 & \text{If Y=+1, a=0} \end{cases}$$

Let $P(Y = 1|x) = p$ $E_Y[C(Y, a = 1)] = P(Y = 1|x) * C(Y = 1, a = 1)) + P(Y = 0|x) * C(Y = 0, a = 1)) = 1 - p$

$E_Y[C(Y, a = 0)] = P(Y = 1|x) * C(Y = 1, a = 0)) + P(Y = 0|x) * C(Y = 0, a = 0)) = 1000p$
We want to label the sample cancerous when $E_Y[C(Y, a = 0)] > E_Y[C(Y, a = 1)]$, i.e. $1000p > 1-p \implies p > 1/1001$.

Problem 2 (Poisson Naive Bayes):

5. Pick the nature of the Naive Bayes model.

   (a) Generative model

(b) Discriminative model
(c) Supervised model
(d) Unsupervised model

Solution:
The correct answers are (a) and (c).
Naive Bayes is a generative supervised model.

6. Let λ be a positive scalar, and assume that $z^{(1)}, z^{(2)}, ...z^{(m)} \in \mathbb{N}$ are m i.i.d observations of a λ-Poisson distributed random variable. Choose 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 $P(Z = k) = \frac{e^{-\lambda}\lambda^k}{k!}$.)

(a) $\frac{\Sigma_{i=1}^m z^{(i)}}{m}$
(b) $\frac{\Sigma_{i=1}^m (z^{(i)})^2}{m}$
(c) $\frac{\Sigma_{i=1}^m \sqrt{z^{(i)}}}{m}$
(d) $\frac{\Sigma_{i=1}^m e^{z^{(i)}}}{m}$

Solution:
The correct answers is (a).

The MLE for a $Poisson(\lambda)$ distribution is the emperical mean.
$P(Z = k) = Poisson(\lambda) = \frac{e^{-\lambda}\lambda^k}{k!}$
$logP(\mathbf{Z}) = -m\lambda + \Sigma_{i=1}^m z^{(i)} log(\lambda) + C$, where C is a constant w.r.t λ. Maximising the log likelihood we get,
$-m + \frac{\Sigma_{i=1}^m z^{(i)}}{\lambda} = 0$
$\implies \lambda_{MLE} = \frac{\Sigma_{i=1}^m z^{(i)}}{m}$

7. Let $D = \{(\mathbf{x^{(1)}}, y^{(1)}), (\mathbf{x^{(2)}}, y^{(2)}), ..., (\mathbf{x^{(n)}}, y^{(n)})\}$, where $\mathbf{x}^{(i)} \in \mathbb{N}^d$ and $y^{(i)} \in 0, 1$ be the training data that you are given. Here, we assume x_i to be the i^{th} dimension of a data point $\mathbf{x}$.

We would like to train a Poisson Naive Bayes classifier, meaning that our model assumes that all the d dimensions of the class conditional distributions, $p(x_i|y)$, are given by independent Poisson distributions. Let $\lambda_0, \lambda_1 \in \mathbb{R}^d$ be the parameters of these Poisson distributions for $y = 0$ and $y = 1$ respectively. Call $p_1 = P(Y = 1)$, and $p_0 = P(Y = 0) = 1 - p_1$. $n_1 = \Sigma_{i=1}^n y_i$ and $n_0 = n - n_1$.

What is the joint distribution $P(\mathbf{x}, y)$?

(a) $p_y \Pi_{j=1}^d Poisson(\lambda_{y,j})$
(b) $p_y \Sigma_{j=1}^d Poisson(\lambda_{y,j})$
(c) $p_y \Pi_{j=1}^n Poisson(\lambda_{y,j})$
(d) $p_y \Sigma_{j=1}^n Poisson(\lambda_{y,j})$

Solution:
The correct answers is (a).

8. We would now like to use MLE to optimise the parameters p_ys and $\lambda_{y,j}$. Pick the true statement regarding this.

(a) p_y and parameters of each component $p(x_i|y)$ CAN be maximised separately. $p_y = \frac{n_y}{n}, \lambda_{y,j} = \frac{\Sigma_{i=1}^n x_j^{(i)} \mathbf{1}_{y_i=y}}{n_y}$.

(b) p_y and parameters of each component $p(x_i|y)$ CANNOT be maximised separately. $p_y = \frac{n_y}{n}, \lambda_{y,j} = \frac{\Sigma_{i=1}^n x_j^{(i)} \mathbf{1}_{y_i=y}}{n_y}$.

(c) p_y and parameters of each component $p(x_i|y)$ CAN be maximised separately. $p_y = \frac{1-n_y}{n}, \lambda_{y,j} = \frac{\Sigma_{i=1}^n x_j^{(i)}}{n_y}$.

(d) p_y and parameters of each component $p(x_i|y)$ CANNOT be maximised separately. $p_y = \frac{n_y}{n}, \lambda_{y,j} = \frac{\Sigma_{i=1}^n x_j^{(i)}}{n_y}$.

Solution:
The correct answers is (a).
$p(y)$ and $p(x_i|y)$ have separate parameters. Hence, they can be maximised separately with respect to their parameters.
n is the total number of datapoints, $n_1 = \Sigma_{i=1}^n y_i$, is the number of times 1 was observed as the label, $n_0 = n - n_1$ is the number of times 0 was observed as the label.
The MLE for $p(y) = Bernoulli(\theta)$ is simply the empirical frequency $p_y = \frac{n_y}{n}$.
Similarly the MLE for a $Poisson(\lambda)$ distribution is just the empirical mean (has been proved in the previous question). Hence we estimate $\lambda_{y,j} = \frac{\Sigma_{i=1}^n x_j^{(i)} \mathbf{1}_{y_i=y}}{n_y}$.

9. Now, we want to use our trained model from Question 8 to minimize the misclassification probability of a new observation, $\mathbf{x} \in \mathcal{X}$,i.e. we predict $y_{pred} = arg\ max_{y \in \mathcal{Y}} P(y|X = x)$. Show that the predicted label y_{pred} for $\mathbf{x}$ is determined by a hyperplane. Choose the correct answer among the following.

   (a) $\mathbf{a} = [log(\frac{\lambda_{1,1}}{\lambda_{0,1}}), ...log(\frac{\lambda_{1,j}}{\lambda_{0,j}})...log(\frac{\lambda_{1,}}{\lambda_{0,d}})]; b = log\frac{p_1}{p_0} + \Sigma_{j=1}^d \lambda_{0,j} - \lambda_{1,j}; y_{pred} = [\mathbf{a}^T\mathbf{x} \geq b]$.

   (b) $\mathbf{a} = [log(\frac{\lambda_{1,1}}{\lambda_{0,1}}), ...log(\frac{\lambda_{1,j}}{\lambda_{0,j}}), ...log(\frac{\lambda_{1,d}}{\lambda_{0,d}})]; b = log\frac{p_0}{p_1} + \Sigma_{j=1}^d \lambda_{1,j} - \lambda_{0,j}; y_{pred} = [\mathbf{a}^T\mathbf{x} \geq b]$.

   (c) $\mathbf{a} = [log(\frac{\lambda_{1,1}}{\lambda_{0,1}}), ..., log(\frac{\lambda_{1,j}}{\lambda_{0,j}}), ..., log(\frac{\lambda_{1,d}}{\lambda_{0,d}})]; b = log\frac{p_1}{p_0} + \Sigma_{j=1}^d \lambda_{0,j} - \lambda_{1,j}; y_{pred} = [\mathbf{a}^T\mathbf{x} \leq b]$.

   (d) $\mathbf{a} = [\frac{\lambda_{1,j}}{\lambda_{0,j}}, ..., \frac{\lambda_{1,j}}{\lambda_{0,j}}, ..., \frac{\lambda_{1,j}}{\lambda_{0,j}}]; b = \frac{p_1}{p_0} + \Sigma_{j=1}^d \lambda_{0,j} - \lambda_{1,j}; y_{pred} = [\mathbf{a}^T\mathbf{x} \geq b]$.

   Solution:
   The correct answers is (b).
   The joined distribution from the Naive Bayes model is:

$$p(x, y) = p_y \Pi_{j=1}^d e^{-\lambda_{y,j}} \frac{\lambda_{y,j}^{x_j}}{x_j!}$$

   We are interested in the decision boundary $p(y = 0|x) = p(y = 1|x)$. We rewrite this as

$$p(y = 0|x) = p(y = 1|x)$$
$$\iff p(x, 0) = p(x, 1)$$
$$\iff p_0 \Pi_{j=1}^d e^{-\lambda_{0,j}} \frac{\lambda_{0,j}^{x_j}}{x_j!} = p_1 \Pi_{j=1}^d e^{-\lambda_{1,j}} \frac{\lambda_{1,j}^{x_j}}{x_j!}$$
$$\iff log(\frac{p_0}{p_1}) + \Sigma_{j=1}^d - \lambda_{0,j} + log(\lambda_{0,j})x_j = \Sigma_{j=1}^d - \lambda_{1,j} + log(\lambda_{1,j})x_j$$

   From the last equation we see that the decision is determined by the hyperpane:

$$0 = log(\frac{p_0}{p_1}) + \Sigma_{j=1}^d \lambda_{1,j} - \lambda_{0,j} + \Sigma_{j=1}^d log \frac{\lambda_{0,j}}{\lambda_{1,j}} x_j$$

$$y_{pred} = [p(y=1|x) \geq p(y=0|x)] = [\mathbf{a}^T\mathbf{x} \geq b]$$

where $\mathbf{a} = \mathbf{a} = [log(\frac{\lambda_{1,1}}{\lambda_{0,1}}), ...log(\frac{\lambda_{1,j}}{\lambda_{0,j}}), ...log(\frac{\lambda_{1,d}}{\lambda_{0,d}})]; b = log\frac{p_0}{p_1} + \Sigma_{j=1}^{d}\lambda_{1,j} - \lambda_{0,j}$

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

    (a) $y_{Bayes} = arg\min_{y\in\mathcal{Y}} E_Y[c(Y,y)|X=x]$
    (b) $y_{Bayes} = arg\min_{y\in\mathcal{Y}} E_Y[c(Y,y)]$
    (c) $y_{Bayes} = arg\min_{y\in\mathcal{Y}} E_Y[c(y,Y)|X=x]$
    (d) $y_{Bayes} = arg\min_{y\in\mathcal{Y}} E_Y[c(y,Y)]$

    Solution:
    The correct answers is (c).
    $y_{Bayes} = arg\min_{y\in\mathcal{Y}} E_Y[c(y,Y)|X=x]$

11. Pick a cost function such that the corresponding decision rule that you have defined in Question 10 for this cost coincides with a decision rule that minimizes the misclassification probability, i.e. $y_{pred} = arg\max_{y\in\mathcal{Y}} P(y|X=x)$.

    (a) $c(y_{pred}, y_{true}) = \mathbf{1}\{y_{pred} \neq y_{true}\}$
    (b) $c(y_{pred}, y_{true}) = (y_{pred} - y_{true})^2$
    (c) $c(y_{pred}, y_{true}) = |y_{pred} - y_{true}|$
    (d) $c(y_{pred}, y_{true}) = \frac{y_{pred}}{y_{true}}$

    Solution:
    The correct answers are (a), (b), and (c).
    The indicted options have a cost of 1 when the label is incorrect, and 0 otherwise. Subsituting the given cost functions in the result from Question 10 gives the misclassification probability.

Problem 3 (Multiclass logistic regression):

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

$$P(y=k|X=\mathbf{x}) = \frac{exp(\mathbf{a_k}^T\mathbf{x})}{\Sigma_j exp(\mathbf{a_j}^T\mathbf{x})}$$

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

$$E(\mathbf{a_1}, \mathbf{a_2}, ..., \mathbf{a_K}) = -ln(\Pi_{n=1}^{N}\Pi_{k=1}^{K}P(y=k|X=\mathbf{x_n})^{t_{nk}}) = -\Sigma_{n=1}^{N}\Sigma_{k=1}^{K}t_{nk}lnP(y=k|X=\mathbf{x_n})$$

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

12. Pick the gradient of the error function with respect to a parameter $\mathbf{a_j}$.

    (a) $\nabla_{\mathbf{a_j}}E(\mathbf{a_1}, \mathbf{a_2}, ..., \mathbf{a_K}) = \Sigma_{n=1}^{N}[P(Y=j|X=\mathbf{x_n}) - t_{nj}]\mathbf{x_n}$
    (b) $\nabla_{\mathbf{a_j}}E(\mathbf{a_1}, \mathbf{a_2}, ..., \mathbf{a_K}) = \Sigma_{n=1}^{N}[P(Y=j|X=\mathbf{x_n}) + t_{nj}]\mathbf{x_n}$
    (c) $\nabla_{\mathbf{a_j}}E(\mathbf{a_1}, \mathbf{a_2}, ..., \mathbf{a_K}) = \Pi_{n=1}^{N}[P(Y=j|X=\mathbf{x_n}) - t_{nj}]\mathbf{x_n}$

(d) $\nabla_{\mathbf{a_j}} E(\mathbf{a_1}, \mathbf{a_2}, ..., \mathbf{a_K}) = \Pi_{n=1}^{N} [P(Y = j|X = \mathbf{x_n}) + t_{nj}]\mathbf{x_n}$

Solution:
The correct answer is (a).

We define $d_k = \mathbf{a_k}^T \mathbf{x}$. The posterior probabilities are given as:

$$P(y = k|X = \mathbf{x}) = \frac{exp(d_k)}{\Sigma_j exp(d_j)} = y_k(\mathbf{x})$$

First, we compute the derivatives of y_k with respect to all d_js:

$$\frac{\partial y_k}{\partial d_j} = y_k(\mathbf{1}_{\{k=j\}} - y_j)$$

This holds because if $j \neq k$, we have:

$$\frac{\partial y_k}{\partial d_j} = \frac{-exp(d_k).exp(d_j)}{[\Sigma_j exp(d_j)]^2} = -y_k.y_j$$

and if $j = k$

$$\frac{\partial y_k}{\partial d_j} = \frac{exp(d_k).\Sigma_j exp(d_j) - exp(d_k).exp(d_k)}{[\Sigma_j exp(d_j)]^2} = y_k.(1 - y_k)$$

Next, we compute the partial derivatives of the summands of $E(...)$

$$\frac{\partial t_{nk} ln y_k(\mathbf{x_n})}{\partial \mathbf{a_j}} = \frac{\partial t_{nk} ln y_k(\mathbf{x_n})}{\partial [y_k(\mathbf{x_n})]} \frac{\partial y_k(\mathbf{x_n})}{\partial d_j} \frac{\partial d_j}{\partial \mathbf{a_j}}$$

where we set $y_{nk} = y_k(\mathbf{x_n})$. We simplify to (using the result $\frac{\partial y_k}{\partial d_j}$ from above):

$$\frac{\partial t_{nk} ln y_k(\mathbf{x_n})}{\partial \mathbf{a_j}} = t_{nk} \frac{1}{y_{nk}} y_{nk}.(\mathbf{1}_{k=j} - y_{nj})\mathbf{x_n} = t_{nk}.(\mathbf{1}_{k=j} - y_{nj})\mathbf{x_n}$$

Then,

$$\begin{aligned}
\nabla_{\mathbf{a_j}} E(...) &= -\Sigma_{n=1}^{N} \Sigma_{k=1}^{K} t_{nk}.(\mathbf{1}_{k=j} - y_{nj})\mathbf{x_n} \\
&= \Sigma_{n=1}^{N} \Sigma_{k=1}^{K} t_{nk} y_{nj} \mathbf{x}_n - \Sigma_{n=1}^{N} \Sigma_{k=1}^{K} t_{nk} \mathbf{1}_{k=j} \mathbf{x_n} \\
&= \Sigma_{n=1}^{N} \left[\Sigma_{k=1}^{K} t_{nk} y_{nj} \right] \mathbf{x}_n - \Sigma_{n=1}^{N} t_{nk} \mathbf{x_n} \\
&= \Sigma_{n=1}^{N} (y_{nj} - t_{nj}) \mathbf{x_n} \\
&= \Sigma_{n=1}^{N} [P(y = j|\mathbf{X} = \mathbf{x_n}) - t_{nj}] \mathbf{x_n}
\end{aligned}$$

(Where we have used the fact that $\Sigma_{k=1}^{K} t_{nk}$ sums to 1 and $y_{nk} = y_k(\mathbf{x_n}) = P(y = k|\mathbf{X} = \mathbf{x_n})$.)