Exercises
Introduction to Machine Learning
SS 2019

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

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

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

Problem 1 (Decision Theory):

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

Answer: Action set: $A = \{Cancerous = 1, Benign = -1\}$

Cost Function:

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

Conditional Distribution: $Bernoulli(y; \sigma(w^T x)$

We would like to have asymmetric costs because we would like to be more sensitive to detecting possibly cancerous samples. We are more concerned with controlling the false negative rate, rather than the false positive rate.

2 Write the action that will minimize the expected cost.

Answer:

$$\begin{aligned} a* &= \mathbb{E}_y[C(Y, a)|x] \\ &= P(Y = 1|x) * C(Y = 1, a = -1)) + P(Y = -1|x) * C(Y = -1, a = 1)) + 0 \\ &= 1000 * P(Y = 1|x) + P(Y = -1|x) \\ &= 1000 * p + (1 - p) \end{aligned}$$

To predict a sample is cancerous, is equivalent to saying that $1000 * p > (1 - p)$. Therefore the optimal action is to label a sample cancerous when $P(Y = 1|x) > 1/999$ and benign all other times.

Problem 2 (Poisson Naive Bayes):

1 Is the Naive Bayes model a generative or a discriminative model? Justify your answer.

Answer: The Naive Bayes model is a generative model because it models the joined data-generating distribution $P(X, Y)$.

2 Let λ be a positive scalar, and assume that $z_1, \ldots, z_m \in \mathbb{N}$ 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 $P(Z = k) = e^{-\lambda}\frac{\lambda^k}{k!}$.)*

3 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.)*

Answer: Let n be the total number of data points, $n_1 = \sum_{i=1}^n y_i$ the number of times '1' was observed, and $n_0 = n - n_1$ the number of '0' accordingly. The Naive Bayes model in our case is

$$p(x, y) = p(y) \prod_{j=1}^{d} p(x_j|y)$$

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 (can easily be checked). Hence we estimate $\lambda_{y,i} = \frac{\sum_{i=1}^n x_{i,j}\mathbf{1}\{y_i=y\}}{n_y}$. The resulting distribution is

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

4 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. $y_{\text{pred}} = \text{argmax}_{y\in\mathcal{Y}} P(y|X = \mathbf{x})$. 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 $\mathbf{a} \in \mathbb{R}^d, b \in \mathbb{R}$.

Answer: The joined distribution from the Naive Bayes model is

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

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

$$\begin{aligned}
& p(y = 0|x) = p(y = 1|x) \\
\iff \quad & p(x, 0) = p(x, 1) \\
\iff \quad & p_0 \prod_{j=1}^{d} e^{-\lambda_{0,j}} \frac{\lambda_{0,j}^{x_j}}{x_j!} = p_1 \prod_{j=1}^{d} e^{-\lambda_{1,j}} \frac{\lambda_{1,j}^{x_j}}{x_j!} \\
\iff & \log\left(\frac{p_0}{p_1}\right) + \sum_{j=1}^{d} -\lambda_{0,j} + \log(\lambda_{0,j})x_j = \sum_{j=1}^{d} -\lambda_{1,j} + \log(\lambda_{1,j})x_j
\end{aligned}$$

From the last equation the claim follows, ie the decision is determined by the hyperplane

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

5 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 $P(X, Y)$.

Answer:

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

6 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., $y_{\text{pred}} = \text{argmax}_{y \in \mathcal{Y}} P(y|X = \mathbf{x})$.
Answer:

$$c(y_{true}, y_{pred}) = \mathbf{1}\{y_{true} \neq y_{pred}\}$$

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 logarithm 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)*:

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

Answer: We define $d_k = \mathbf{a}_k^\top \mathbf{x}$. Then the posterior probabilities are given as

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

First, we compute the derivatives of y_k with respect to all d_j's:

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

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

$$\frac{\partial y_k}{\partial d_j} = \frac{-\exp(d_k) \cdot \exp(d_j)}{\left[\sum_j \exp(d_j)\right]^2} = -y_k \cdot y_j$$

and if $j = k$:

$$\frac{\partial y_k}{\partial d_j} = \frac{\exp(d_k) \cdot \sum_j \exp(d_j) - \exp(d_k) \cdot \exp(d_k)}{\left[\sum_j \exp(d_j)\right]^2} = y_k(1 - y_k)$$

Next, we compute the partial derivative of the summands of $E(\dots)$

$$\frac{\partial t_{nk} \ln\, y_k(\mathbf{x}_n)}{\partial \mathbf{a}_j} = \frac{\partial t_{nk} \ln\, y_k(\mathbf{x}_n)}{\partial\, [y_k(\mathbf{x}_n)]} \frac{\partial y_k(\mathbf{x}_n)}{\partial d_j} \frac{\partial d_j}{\partial \mathbf{a}_j} = \frac{\partial t_{nk} \ln\, y_{nk}}{\partial y_{nk}} \frac{\partial y_{nk}}{\partial d_j} \frac{\partial d_j}{\partial \mathbf{a}_j}$$

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

$$\frac{\partial t_{nk} \ln\, y_k(\mathbf{x}_n)}{\partial \mathbf{a}_j} = t_{nk} \frac{1}{y_{nk}} y_{nk} \cdot (\mathbf{1}_{\{k=j\}} - y_{nj}) \cdot \mathbf{x}_n = t_{nk}(\mathbf{1}_{\{k=j\}} - y_{nj})\mathbf{x}_n$$

Then,

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

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