Machine Learning Homework Sheet 04

Linear Classification

1 Linear classification

Problem 1: We want to create a generative binary classification model for classifying *nonnegative* one-dimensional data. This means, that the labels are binary ($y \in \{0, 1\}$) and the samples are $x \in [0, \infty)$.

We place a uniform prior on y

$$p(y = 0) = p(y = 1) = \frac{1}{2}.$$

As our samples x are nonnegative, we use exponential distributions (and not Gaussians) as class conditionals:

$$p(x \mid y = 0) = \text{Expo}(x \mid \lambda_0) \qquad \text{and} \qquad p(x \mid y = 1) = \text{Expo}(x \mid \lambda_1),$$

where $\lambda_0 \neq \lambda_1$. Assume, that the parameters λ_0 and λ_1 are known and fixed.

a) What is the name of the posterior distribution $p(y \mid x)$? You only need to provide the name of the distribution (e.g., "normal", "gamma", etc.), not estimate its parameters.

Bernoulli.

Remark: y can only take values in $\{0, 1\}$, so obviously Bernoulli is the only possible answer.

b) What values of x are classified as class 1?
(As usual, we assume that the classification decision is $y_{predicted} = \arg\max_k p(y = k \mid x)$)

Sample x is classified as class 1 if $p(y = 1 \mid x) > p(y = 0 \mid x)$. This is the same as saying

$$\frac{p(y = 1 \mid x)}{p(y = 0 \mid x)} \overset{!}{>} 1 \quad \text{or equivalently} \quad \log \frac{p(y = 1 \mid x)}{p(y = 0 \mid x)} \overset{!}{>} 0.$$

$$\begin{aligned}
\log \frac{p(y = 1 \mid x)}{p(y = 0 \mid x)} &= \log \frac{p(x \mid y = 1)p(y = 1)}{p(x \mid y = 0)p(y = 0)} \\
&= \log \frac{p(x \mid y = 1)}{p(x \mid y = 0)} \\
&= \log \frac{\lambda_1 \exp(-\lambda_1 x)}{\lambda_0 \exp(-\lambda_0 x)} \\
&= \log \frac{\lambda_1}{\lambda_0} + \lambda_0 x - \lambda_1 x = \log \frac{\lambda_1}{\lambda_0} + (\lambda_0 - \lambda_1)x
\end{aligned}$$

Upload a single PDF file with your solution to Moodle by 18.11.2018, 23:59 CET. We recommend to typeset your solution (using LaTeX or Word), but handwritten solutions are also accepted.
If your handwritten solution is illegible, it won't be graded and you waive your right to dispute that.

The inequality that we are interested in solving is

$$(\lambda_0 - \lambda_1)x + \log\frac{\lambda_1}{\lambda_0} > 0$$
$$(\lambda_0 - \lambda_1)x > \log\frac{\lambda_0}{\lambda_1}$$

We have to be careful, because if $(\lambda_0 - \lambda_1) < 0$, diving by it will flip the inequality sign. Hence the answer is

$$\begin{cases} x \in \left(\frac{\log\lambda_0 - \log\lambda_1}{\lambda_0 - \lambda_1}, \infty\right) & \text{if} \quad \lambda_0 > \lambda_1; \\ x \in \left[0, \frac{\log\lambda_0 - \log\lambda_1}{\lambda_0 - \lambda_1}\right) & \text{if} \quad \lambda_0 < \lambda_1. \end{cases}$$

Problem 2: Assume you have a linearly separable data set. What properties does the maximum likelihood solution for the decision boundary $\boldsymbol{w}$ of a logistic regression model have? Assume that $\boldsymbol{w}$ includes the bias term.

What is the problem here and how do we prevent it?

Finding a separating hyperplane puts all training points on the correct side (and thus ensures posterior class probabilities > 0.5 for each training point). Taking the magnitude of $\boldsymbol{w}$ to infinity then assigns each training point a posterior class probability of 1. This happens because of the shape of the logistic sigmoid function—it becomes a heaviside (step) function in this particular case.

We can prevent this by adding a weight regularization term to the loss function, which will make the loss function bounded.

Problem 3: Show that the softmax function is equivalent to a sigmoid in the 2-class case.

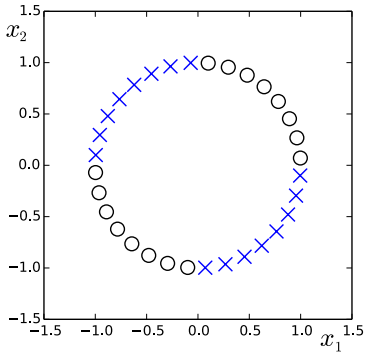
$$\begin{aligned} \frac{\exp(\boldsymbol{w}_1^T\boldsymbol{x})}{\exp(\boldsymbol{w}_1^T\boldsymbol{x}) + \exp(\boldsymbol{w}_0^T\boldsymbol{x})} &= \frac{1}{1 + \exp(\boldsymbol{w}_0^T\boldsymbol{x})/\exp(\boldsymbol{w}_1^T\boldsymbol{x})} \\ &= \frac{1}{1 + \exp(\boldsymbol{w}_0^T\boldsymbol{x} - \boldsymbol{w}_1^T\boldsymbol{x})} \\ &= \frac{1}{1 + \exp(-(\boldsymbol{w}_1 - \boldsymbol{w}_0)^T\boldsymbol{x})} \\ &= \sigma(\hat{\boldsymbol{w}}^T\boldsymbol{x}) \end{aligned}$$

where $\hat{\boldsymbol{w}} = \boldsymbol{w}_1 - \boldsymbol{w}_0$.

One conclusion we can draw from this is that if we have C parameter vectors $\boldsymbol{w}_c$ for C classes, the logistic regression model is unidentifiable. This means that adding the same constant $k \in \mathbb{R}$ to both vectors $\boldsymbol{w}_i := \boldsymbol{w}_i + k$ for $i \in \{0, 1\}$ would lead to the same logistic regression model. We can fix this

issue by adding a constraint $\boldsymbol{w}_0 = \mathbf{0}$, which is what is done implicitly when we use sigmoid (instead of 2-class softmax) in binary classification.

Problem 4: Which basis function $\phi(x_1, x_2)$ makes the data in the example below linearly separable (crosses in one class, circles in the other)?



There are many choices of basis functions to achieve this goal. One example is

$$\phi(x_1, x_2) = x_1 x_2$$