Exercises
Learning and Intelligent Systems
SS 2017

Series 5, May 22nd, 2016
(Probabilistic Modeling & Autoencoders)

It is not mandatory to submit solutions and sample solutions will be published in two weeks. If you choose to submit your solution, please send an e-mail from your ethz.ch address with subject Exercise5 containing a PDF (\LaTeX or scan) to natalie.davidson@inf.ethz.ch until Monday, May 22th 2017.

Problem 1 (Independence Assumptions of Naive Bayes Classifiers):
Consider a naive Bayes classifier with binary class variable $C \in \{0, 1\}$ and two binary features $X_1 \in \{0, 1\}$ and $X_2 \in \{0, 1\}$. Assume that $X_1$ and $X_2$ are truly independent. You are given the following probabilities:

$$
\begin{align*}
P(X_1 = 1 | C = 1) &= p \\
P(X_1 = 1 | C = 0) &= 1 - p \\
P(X_2 = 0 | C = 1) &= q \\
P(X_2 = 0 | C = 0) &= 1 - q \\
P(C = 0) &= P(C = 1) = 0.5
\end{align*}
$$

(a) Given a test sample with $X_1 = 1$ and $X_2 = 0$, compute the decision rule for classifying the example as belonging to class 1 in terms of $q$ and $p$. Reformulate the decision rule in the form $p \geq \ldots$.

(b) We extend the naive Bayes classifier by adding another feature $X_3$ which is simply a copy of $X_2$. Again, compute the decision rule of the classifier in terms of $q$ and $p$. Reformulate the decision rule in the form $p \geq \ldots$.

(c) Compare the decision boundaries of (a) and (b) by varying the value of $q$ between 0 and 1. Show where the second rule makes mistakes relative to the first (correct) decision rule.

Problem 2 (Bayesian optimal decisions for logistic regression):
Apply Bayesian decision theory to derive the optimal decision rule for logistic regression in the following setting:

- Estimated conditional distribution: $\hat{P}(y | x) = \begin{cases} 
\sigma(w^T x) & \text{if } y = 1 \\
1 - \sigma(w^T x) & \text{if } y = -1
\end{cases}$

- Action set: $\{+1, -1, D\}$

- Cost function: $C(y, a) = \begin{cases} 
1[y \neq a] & \text{if } a \in \{+1, -1\} \\
c < 0.5 & \text{if } a = D
\end{cases}$

Here, $1[\cdot]$ denotes the indicator function.
Problem 3 (Bayesian optimal decisions for regression with asymmetric costs):

Apply Bayesian decision theory to derive the optimal decision rule for linear regression in the following setting:

- Estimated conditional distribution: \( \hat{P}(y|x) = N(y; w^T x, \sigma^2) \)
- Action set: \( \mathbb{R} \)
- Cost function: \( C(y, a) = c_1 \max(y - a, 0) + c_2 \max(a - y, 0) \)

Here, \( c_1 \) and \( c_2 \) denote positive real valued constants.

Problem 4 (Optional) - (Autoencoders and PCA):

In this exercise, we analyze dimensionality reduction using autoencoders with linear activation functions and relate them to principal component analysis (PCA). We consider the following setup: let \( D = \{x_1, \ldots, x_N\} \) be given inputs, with \( x_i \in \mathbb{R}^n \). Let \( X = [x_1, \ldots, x_N] \in \mathbb{R}^{n \times N} \) be the matrix formed from the inputs. Assume that we compute \( p \) hidden activations for every input \( x_i \) example according to \( h_i = \phi_1(W_1 x_i + b_1) \), where \( \phi_1(\cdot) \) is an activation function applied element-wise, \( W_1 \in \mathbb{R}^{p \times n} \) are the input weights, and \( b_1 \in \mathbb{R}^p \) are biases. Note that we can express the computation of all hidden activations as \( H = \phi_1(W_1 X + b_1 u^T) \), where \( u \) is a vector containing only ones of size \( N \). For this analysis, assume that \( \phi_1(x) = x \). The weights and biases of the autoencoder are selected as

\[
\arg \min_{W_1, W_2, b_1, b_2} \|X - Y\|^2. \tag{1}
\]

(a) Consider the squared-error criterion given the hidden activations, i.e. \( \|X - (W_2 H + b_2 u^T)\|^2 \). Derive an expression for the biases \( b_2 \) in terms of \( X, H \) and \( W_2 \). Substitute your expression into the error and rewrite it in the form \( \|X' - W_2 H'\|^2 \), where \( X' \) (\( H' \)) depends only on \( X \) (\( H \)) and constants.

(b) Compare the problem of minimizing \( \|X' - W_2 H'\|^2 \) with the problem of computing the PCA from the lecture. Read off the optimal \( W_2 \) and \( H' \). They should be expressed up to an arbitrary non-singular linear transform given by a \( p \times p \) matrix \( T \).

(c) Show that the obtained solution for \( H' \) can actually be generated by proper choices of \( W_1 \) and \( b_1 \).

(d) Comment on the relation of \( W_1 \) to \( W_2 \).

(e) Comment on the transformation of the input computed by the autoencoder with respect to PCA.