Exercises
Learning and Intelligent Systems
SS 2016

LAS Group, Institute for Machine Learning
Dept. of Computer Science, ETH Zürich
Prof. Dr. Andreas Krause
Web: `http://las.inf.ethz.ch/teaching/lis-s16/`
Email questions to:
Felix Berkenkamp, `befelix@inf.ethz.ch`

Series 5, May 3rd, 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 `lis2015@lists.inf.ethz.ch` until Monday, May 16th 2016.

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:

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

(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.

Solution 1 (Independence Assumptions of Naive Bayes Classifiers):

(a) Exploiting the independence assumption we compute

$$\begin{aligned} P(X_1 = 1, X_2 = 0, C = 1) &= P(X_1 = 1, X_2 = 0|C = 1)P(C = 1) && (1) \\ &= P(X_1 = 1|C = 1)P(X_2 = 0|C = 1)P(C = 1) && (2) \\ &= pq\frac{1}{2}, && (3) \end{aligned}$$

and, similarly,

$$\begin{aligned} P(X_1 = 1, X_2 = 0, C = 0) &= P(X_1 = 1|C = 0)P(X_2 = 0|C = 0)P(C = 0) && (4) \\ &= (1 - p)(1 - q)\frac{1}{2}. && (5) \end{aligned}$$

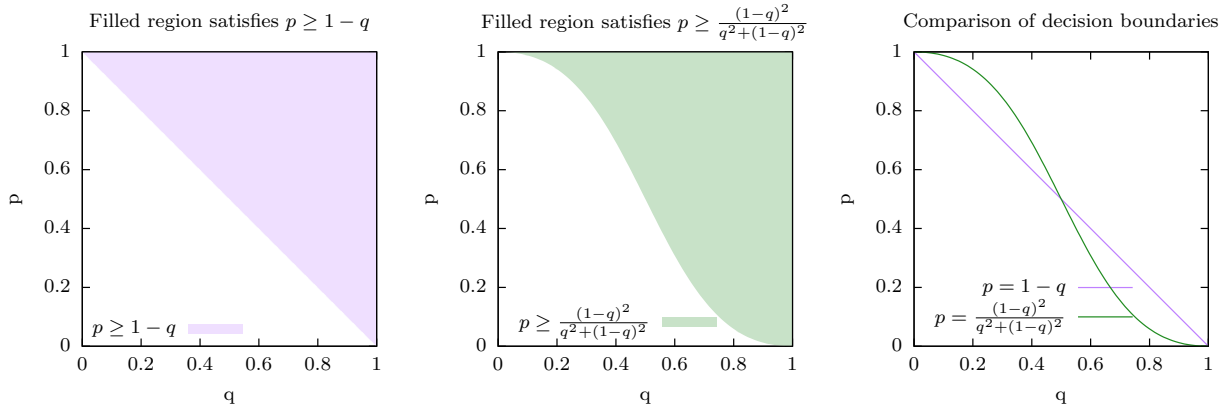

Figure 1: Decision boundaries

Consequently, we want

$$pq\frac{1}{2} \geq (1-p)(1-q)\frac{1}{2}. \tag{6}$$

Solving for p yields

$$p \geq 1 - q. \tag{7}$$

(b) Similarly, using the replicated feature one obtains

$$pqq\frac{1}{2} \geq (1-p)(1-q)(1-q)\frac{1}{2} \tag{8}$$

Solving for p yields

$$p \geq \frac{(1-q)^2}{q^2+(1-q)^2}. \tag{9}$$

(c) The two boundaries are shown in Figure 1. Furthermore, the figure shows where one decision rule differs relative to the other (this is the region between the purple and the green line in the rightmost figure).

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: $\hat{P}(y|\mathbf{x}) = \begin{cases} \sigma(\mathbf{w}^T\mathbf{x}) & \text{if } y = 1 \\ 1 - \sigma(\mathbf{w}^T\mathbf{x}) & \text{if } y = -1 \end{cases}$
- Action set: $\{+1, -1, D\}$
- Cost function: $C(y, a) = \begin{cases} \mathbf{1}[y \neq a] & \text{if } a \in \{+1, -1\} \\ c < 0.5 & \text{if } a = D \end{cases}$

Here, $\mathbf{1}[\cdot]$ denotes the indicator function.

Solution 2 (Bayesian optimal decisions for logistic regression):

In the Bayesian optimal decision framework we want to pick the action which minimizes the expected loss:

$$a^\star = \underset{a \in \mathcal{A}}{\operatorname{argmin}}\, \mathbb{E}_y[C(y, a) \mid x] = \underset{a \in \mathcal{A}}{\operatorname{argmin}} \int C(y, a) p(y \mid x) dy$$

In our case

$$a^\star = \underset{a \in \mathcal{A}}{\operatorname{argmin}} \sum_y C(y, a) p(y \mid x) = \underset{a \in \mathcal{A}}{\operatorname{argmin}} \sum_y \left(\mathbf{1}[y \neq a]\mathbf{1}[D \neq a] + c\mathbf{1}[D = a]\right) p(y \mid x)$$

Let's analyze the expected cost of each action separately:

$$\mathbb{E}[C(y, a) \mid x] = \begin{cases} p(y = -1 \mid x) & \text{if } a = 1 \\ p(y = 1 \mid x) & \text{if } a = -1 \\ c & \text{if } a = D \end{cases}$$

We can conclude that we should only pick an action $a \in \{-1, 1\}$ if the probability of action $-a$ being correct is lower than than c!

$$a^\star = \begin{cases} y & \hat{P}(-y \mid x) < c \\ D & \text{otherwise} \end{cases}$$

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: $\hat{P}(y|\mathbf{x}) = \mathcal{N}(y; \mathbf{w}^T\mathbf{x}, \sigma^2)$
- 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.

Solution 3 (Bayesian optimal decisions for regression with asymmetric costs):

We aim to minimize the expected risk $\mathbb{E}[C(y, a)] = \int C(y, a)\hat{P}(y|\mathbf{x})\mathrm{d}y$. Differentiating with respect to a yields

$$\frac{\partial}{\partial a}\mathbb{E}[C(y, a)] = \frac{\partial}{\partial a}\int_{-\infty}^{\infty} C(y, a)\hat{P}(y|\mathbf{x})\mathrm{d}y \tag{10}$$

$$= \frac{\partial}{\partial a}\int_{-\infty}^{\infty} c_1 \max(y - a, 0)\hat{P}(y|\mathbf{x})\mathrm{d}y + \int_{-\infty}^{\infty} c_2 \max(a - y, 0)\hat{P}(y|\mathbf{x})\mathrm{d}y \tag{11}$$

$$= -c_1\int_{a}^{\infty} \hat{P}(y|\mathbf{x})\mathrm{d}y + c_2\int_{-\infty}^{a} \hat{P}(y|\mathbf{x})\mathrm{d}y \tag{12}$$

$$= -c_1[1 - \Phi(a; \mathbf{w}^T\mathbf{x}, \sigma^2)] + c_2\Phi(a; \mathbf{w}^T\mathbf{x}, \sigma^2), \tag{13}$$

where $\Phi(u; v, w)$ is the CDF of the normal distribution with mean v and variance w. Setting the derivative to zero and applying some algebraic manipulations yields

$$\Phi(a; \mathbf{w}^T\mathbf{x}, \sigma^2) = \frac{c_1}{c_1 + c_2}. \tag{14}$$

Noting that $\Phi(u; v, w) = \Phi((u - v)/\sqrt{w}; 0, 1)$, we can rewrite the above equation as

$$\Phi\left(\frac{a - \mathbf{w}^T\mathbf{x}}{\sigma}; 0, 1\right) = \frac{c_1}{c_1 + c_2}. \tag{15}$$

Applying the inverse CDF of the standard normal distribution Φ^{-1}, and rewriting the result in terms of a, yields the optimal action

$$a^* = \mathbf{w}^T\mathbf{x} + \sigma\Phi^{-1}\left(\frac{c_1}{c_1 + c_2}\right). \tag{16}$$

Thus, if for example the cost of underestimation c_1 is smaller than the cost of overestimation c_2, i.e. $c_1 < c_2$, then $\Phi^{-1}(\frac{c_1}{c_1+c_2}) < 0$. In this way the risk of overestimation is reduced.

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 $D = \{x_1, \dots, x_N\}$ be given inputs, with $x_i \in \mathbb{R}^n$. Let $X = [x_1, \dots, x_N] \in \mathbb{R}^{n \times N}$ 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 $\phi_1(x) = x$. Given the hidden activations H and output weights $W_2 \in \mathbb{R}^{n \times p}$ as well as biases $b_2 \in \mathbb{R}^n$, the output of the autoencoder is computed as $Y = \phi_2(W_2 H + b_2 u^T)$, where again we assume $\phi_2(x) = x$. The weights and biases of the autoencoder are selected as

$$\arg \min_{W_1, W_2, b_1, b_2} \|X - Y\|^2. \tag{17}$$

(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.

Solution 4 (Autoencoders and PCA):

This solution is closely based on *H. Bourlard, and Y. Kamp,* Auto-association by multilayer perceptrons and singular value decomposition, *Biological Cybernetics, Springer-Verlag, vol. 59, pp. 291–294 1988.*

(a) Note that $\|A\|^2 = \mathrm{tr}(AA^T)$, where $\mathrm{tr}(B)$ denotes the trace of matrix B. Furthermore, note that $\mathrm{tr}(\cdot)$ is a linear operator and that $\mathrm{tr}(A) = \mathrm{tr}(A^T)$. Hence,

$$\begin{aligned}
\|X - (W_2 H + b_2 u^T)\|^2 &= \mathrm{tr}((X - (W_2 H + b_2 u^T))(X - (W_2 H + b_2 u^T))^T) && (18)\\
&= \mathrm{tr}(XX^T) - 2\mathrm{tr}(X(W_2 H + b_2 u^T)^T) + \mathrm{tr}((W_2 H + b_2 u^T)(W_2 H + b_2 u^T)^T) && (19)\\
&= \mathrm{tr}(XX^T) - 2\mathrm{tr}(XH^T W_2) - 2\mathrm{tr}(X u b_2^T) + \mathrm{tr}(W_2 H H^T W_2) && (20)\\
&\quad + 2\mathrm{tr}(W_2 H u b_2^T) + \mathrm{tr}(b_2 u^T u b_2^T) && (21)
\end{aligned}$$

Dropping all terms that do not depend on b_2 (these will cancel once we compute the gradient with respect to b_2) and noting that $u^T u = N$, we get

$$-2\mathrm{tr}(X u b_2^T) + 2\mathrm{tr}(W_2 H u b_2^T) + N\mathrm{tr}(b_2 b_2^T). \tag{22}$$

Computing the gradient of the above expression with respect to b_2 yields

$$\begin{aligned}
\nabla_{b_2}[-2\mathrm{tr}(X u b_2^T) + 2\mathrm{tr}(W_2 H u b_2^T) + N\mathrm{tr}(b_2 b_2^T)] &= -2\nabla_{b_2}\mathrm{tr}(X u b_2^T) + 2\nabla_{b_2}\mathrm{tr}(W_2 H u b_2^T) + N\nabla_{b_2}\mathrm{tr}(b_2 b_2^T) && (23)\\
&= -2Xu + 2W_2 H u + 2N b_2. && (24)
\end{aligned}$$

Equating to zero and solving for b_2 yields

$$b_2 = \frac{1}{N}(X - W_2 H)u. \tag{25}$$

Substituting this result into $\|X - (W_2 H + b_2 u^T)\|^2$ gives

$$\|X - (W_2 H + b_2 u^T)\|^2 = \|X - (W_2 H + \frac{1}{N}(X - W_2 H)uu^T)\|^2 \tag{26}$$

$$= \|X(I - \frac{1}{N}uu^T) - W_2[H(I - \frac{1}{N}uu^T)]\|^2 \tag{27}$$

$$= \|X' - W_2 H'\|^2, \tag{28}$$

where $X' = X(I - \frac{1}{N}uu^T)$ and $H' = H(I - \frac{1}{N}uu^T)$.

(b) Assume that $p < n$, i.e. W_2 has rank p. From the lecture we know that the optimal solution to $\|X' - W_2 H'\|^2$ is given when the product $W_2 H$ is the best rank p approximation of X'. This approximation can be obtained by computing the singular value decomposition (SVD) of X', i.e.

$$X' = U\Sigma V^T, \tag{29}$$

where U and V are matrices of size $n \times n$ and $N \times n$, respectively, consisting of the normalized eigenvectors of $X'X'^T$ and $X'^T X'$, respectively. Σ is a diagonal matrix with entries $\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}$ on the diagonal, where λ_i is the i-th largest eigenvalue (assume the eigenvalues to be sorted such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$). In this setting, the best rank p approximation of X' is given by

$$W_2 H' = \tilde{U}\tilde{\Sigma}\tilde{V}^T, \tag{30}$$

where $\tilde{\Sigma}$ is a diagonal matrix with the p largest eigenvalues $\lambda_1, \ldots, \lambda_p$ as its entries, and where $\tilde{U}$ and $\tilde{V}^T$ are formed by the first p columns of U and V, respectively.

Consequently, denoting by T an arbitrary invertible $p \times p$ matrix, we have $W_2 = \tilde{U}T^{-1}$ and $H' = T\tilde{\Sigma}\tilde{V}^T$.

(c) It remains to show that the H' can actually be generated by proper choices of W_1 and b_1. The output H of the first layer is computed according to $H = W_1 X + b_1 u^T$. Multiplying from the right both sides by $(I - \frac{1}{N}uu^T)$ gives

$$H(I - \frac{1}{N}uu^T) = H' = W_1 X' + b_1 u^T (I - \frac{1}{N}uu^T). \tag{31}$$

Hence, we need that

$$T\tilde{\Sigma}\tilde{V}^T = H' \tag{32}$$

$$= W_1 X' + b_1 u^T (I - \frac{1}{N}uu^T). \tag{33}$$

Observe that

$$b_1 u^T (I - \frac{1}{N}uu^T) = b_1 u^T - \frac{1}{N} b_1 u^T uu^T \tag{34}$$

$$= b_1 u^T - \frac{1}{N} b_1 N u^T \tag{35}$$

$$= b_1 u^T - b_1 u^T \tag{36}$$

$$= 0, \tag{37}$$

and hence w_1 is arbitrary. Furthermore,

$$W_1 X' = T\tilde{\Sigma}\tilde{V}^T \tag{38}$$

$$= T\tilde{U}^T X', \tag{39}$$

where we used $X' = U\Sigma V^T$. Consequently, $W_1 = T\tilde{U}^T$ is a proper choice for W_1.

(d) We found that $W_1 = T\tilde{U}^T$ and $W_2 = \tilde{U}T^{-1}$. Setting $T = I$, we have $W_2 = W_1^T$.

(e) An auto-encoder essentially computes the PCA (up to an invertible linear transform).