Decision-theoretic machine learning

List of questions

In order to pass and get 4.0 grade, you need to correctly solve *one* of the questions from the list below. To get 5.0 grade, you need to solve *two* questions. The solution(s) should be sent (in PDF format, do not send Word documents!) to `wkotlowski@cs.put.poznan.pl` with a term '`[DTML]`' in the title. The deadline is **30th June, 2018**.

List of questions

1. Consider the *absolute value loss function* defined as:

$$\ell(y, \widehat{y}) = |y - \widehat{y}|.$$

   Show that if y is generated from some distribution $P(y)$, then the Bayes optimal decision y^*, i.e., the one minimizing the expected loss:

$$y^* \;=\; \arg\min_{\widehat{y}} \mathbb{E}_{y \sim P(y)} \left[\ell(y, \widehat{y}) \right],$$

   is the *median* of distribution P, i.e. $y^* \;=\; \mathrm{median}(y)$.

2. In binary classification with the zero-one loss function, the Bayes (optimal) classifier is given by:

$$h^*(\boldsymbol{x}) = \mathrm{sgn}(\eta(\boldsymbol{x}) - 1/2), \qquad \text{where} \;\; \eta(\boldsymbol{x}) = P(y = 1|\boldsymbol{x}).$$

   Derive the Bayes classifier for a loss function with classification costs (*cost-sensitive loss function*):

$$\ell(y, \widehat{y}) = \begin{cases} 0 & \text{if } \; y = \widehat{y}, \\ 1 & \text{if } \; y = 1, \, \widehat{y} = -1, \\ \beta & \text{if } \; y = -1, \, \widehat{y} = 1. \end{cases}$$

   Note: if $\beta = 1$, we get a standard zero-one loss; in this case the derived Bayes classifier should agree with the Bayes classifier for the zero-one loss.

3. *Naive Bayes* classifier is based on the assumption that features are independent in a given class, i.e. for any class index k and any $\boldsymbol{x} = (x_1, \ldots, x_m)$,

$$P(\boldsymbol{x}|y = k) = \prod_{j=1}^{m} P(x_j|y = k) \,.$$

Does this assumption imply (or is implied by) the assumption that features are *unconditionally* independent, i.e.:

$$P(\boldsymbol{x}) = \prod_{j=1}^{m} P(x_j) \,.$$

Justify your answer by either giving a counter-example (if the answer is *no*) or providing a proof (if the answer is *yes*). *Note*: you need to answer two questions here: whether the first assumption implies the second, and whether the second assumption implied the first.

4. Is the *Naive Bayes* classifier is a linear classifier, i.e., whether it corresponds to a classification function:

$$h(\boldsymbol{x}) = \operatorname{sgn}(f(\boldsymbol{x})), \qquad \text{where} \;\; f(\boldsymbol{x}) = w_0 + \sum_{j=1}^{m} w_j x_j?$$

Justify your answer by providing explicit calculations. For simplicity, restrict the answer to the case of binary features, i.e. when $x \in \{0, 1\}$.

5. The optimal solution to the linear regression problem is given by:

$$\widehat{\boldsymbol{w}} = \left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top\right)^{-1} \left(\sum_{i=1}^{n} y_i \boldsymbol{x}_i\right).$$

What happens if the number of features m is *larger* than the number of training examples n? Justify your answer. Furthermore, propose a way to cope with this problem.

6. Show that minimization of the zero-one loss within the class of linear classifiers is *NP-hard* (propose a polynomial reduction to another NP-hard problem).

7. Show that the loss functions below:

   - squared loss: $\ell(f) = (1 - f)^2$,
   - logistic loss: $\ell(f) = \log\left(1 + e^{-f}\right)$,
   - hinge loss: $\ell(f) = \max\{0, 1 - f\}$,
   - exponential loss: $\ell(f) = e^{-f}$.

are *convex* as functions of the margin f.

8. Show that if training examples $(\boldsymbol{x}, y)$ are generated by first drawing a label $y \in \{-1, 1\}$ from some distribution $P(y)$ and then drawing $\boldsymbol{x}|y \sim N(\mu_y, \Sigma)$ (i.e., each class has its own mean vector, but the covariance matrix is shared between classes), then $\log \frac{\eta(\boldsymbol{x})}{1-\eta(\boldsymbol{x})}$ is a linear function of $\boldsymbol{x}$, where $\eta(\boldsymbol{x}) = P(y = 1|\boldsymbol{x})$. For simplicity, you can assume that Σ is an identity matrix.

9. Prove that all loss functions below are classification calibrated. Furthermore, derive the Bayes classifier for each loss:

   - square loss: $\ell(f) = (1 - f)^2$,
   - logistic loss: $\ell(f) = \log\left(1 + e^{-f}\right)$,
   - hinge loss: $\ell(f) = \max\{0, 1 - f\}$,
   - exponential loss: $\ell(f) = e^{-f}$.

10. Prove that the class of rectangles on a plane has Vapnik-Chervonenkis dimension exactly equal to 4.

11. Prove that structured support vector machines with Hamming loss as the task loss and the scoring function of the following form:

$$f(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{m} f_i(\boldsymbol{x}, y_i)$$

    boil down to binary relevance with binary support vector machines.

12. Prove that conditional random fields with the scoring function of the following form:

$$f(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{m} f_i(\boldsymbol{x}, y_i)$$

    boil down to binary relevance with logistic regression.