Decision-theoretic machine learning

List of problems

In order to pass the course you need to solve some of the problems described below from 4 different topics. For each solved problem you can get max. 1 point. However, you cannot get more than 1 point from a given topic. The final mark will be given according to the following rule:

- 3.0 points – 5.0
- 2.0 points – 4.0
- 1.0 points – 3.0

Your solutions should be sent (in a LaTeX-generated PDF file) to both instructors via email. Please use tag `[DTML]` in the title. The deadline is **June 30, 2019**.

List of questions

Statistical learning theory

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. 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).

4. 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.

5. 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.

6. 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: $\ell(f) = e^{-f}$.

7. Prove that the class of linear functions in $\mathbb{R}^n$ has the Vapnik-Chervonenkis dimension equal to $n+1$.
   Note: To prove that you need to show that there are some $n+1$ points that can be shattered (the simpler part) and there does not exist a set of $n+2$ points that can be shattered (the harder part). The knowledge about linear algebra will be very useful.

Learning algorithms

1. Derive the splitting criterion of the decision tree growing algorithm for the squared-error loss and 0/1 loss. Show that the value of the splitting criterion can be updated incrementally when scanning examples one by one in the sorted order of values on a given variable.

2. *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.

3. 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 $x \in \{0, 1\}$.

4. 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.

5. For the polynomial kernel of degree $p = 2$ given by the following equation:

$$K_2(\boldsymbol{x}_i, \boldsymbol{x}_j) = (1 + \boldsymbol{x}_i^\top \boldsymbol{x}_j)^2$$

derive the corresponding primal form of the feature map.

Bipartite ranking

1. Let $f(\boldsymbol{x})$ be a ranker which assigns to each learning example $\boldsymbol{x}$ a random number taken uniformly from $[0, 1]$. Given the prior probability of positive class is equal to $p = P(y = 1)$, compute the rank risk of such f as well as its zero-one risk.

2. It is shown on the slides that given a fixed empirical zero-one risk $\widehat{L}_{0/1}(f)$ of f, f can have arbitrary bad (arbitrarily close to 1) empirical rank risk $\widehat{L}_{\mathrm{rnk}}(f)$. Try to find the opposite bound: assume that the empirical rank risk $\widehat{L}_{\mathrm{rnk}}(f)$ of f is given and equal to x. What is the largest possible empirical zero-one risk of f? Write down the answer in terms of x and $p = P(y = 1)$ (the prior probability of positive class).

3. Reduction from ranking to pairwise binary classification presented on the slides:

   - What would happen if we did not impose a structure on the function space $\tilde{f}(\tilde{\boldsymbol{x}}_k) = f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j)$ in the original reduction?
   - What would happen in the reduction if we included the negative examples as well? In other words, what would change if the data transformation would produce a data set of the form: for any pair (i, j) with $i \neq j$:

   $$\tilde{\boldsymbol{x}}_k = (\boldsymbol{x}_i, \boldsymbol{x}_j), \qquad \tilde{y}_k = \mathrm{sgn}(y_i - y_j)$$

   Note: these two questions are independent of each other.

4. Define:

   $$K(\boldsymbol{x}, \boldsymbol{x}') = \eta(\boldsymbol{x})(1 - \eta(\boldsymbol{x}'))\Big([\![f(\boldsymbol{x}) < f(\boldsymbol{x}')]\!] + \frac{1}{2} [\![f(\boldsymbol{x}) = f(\boldsymbol{x}')]\!] \Big),$$

   where $\eta(\boldsymbol{x}) = P(y = 1|\boldsymbol{x})$. Show that the ranking risk can be rewritten as:

   $$\begin{aligned} L_{\mathrm{rnk}}(f) &= \frac{1}{p(1-p)} \mathbb{E}_{(\boldsymbol{x},\boldsymbol{x}')} \left[K(\boldsymbol{x}, \boldsymbol{x}') \right] \\ &= \frac{1}{2p(1-p)} \mathbb{E}_{(\boldsymbol{x},\boldsymbol{x}')} \left[K(\boldsymbol{x}, \boldsymbol{x}') + K(\boldsymbol{x}', \boldsymbol{x}) \right]. \end{aligned}$$

   where $p = P(y = 1)$ is the prior probability of positive class

5. Based on the result of the previous question, argue that the Bayes ranker $f^*(\boldsymbol{x})$ minimizes $K(\boldsymbol{x}, \boldsymbol{x}') + K(\boldsymbol{x}', \boldsymbol{x})$ for every $(\boldsymbol{x}, \boldsymbol{x}')$. Show that this implies:

   $$f^*(\boldsymbol{x}) > f^*(\boldsymbol{x}') \qquad \text{if and only if} \qquad \eta(\boldsymbol{x}) > \eta(\boldsymbol{x}'),$$

   i.e., the Bayes ranker $f^*(\boldsymbol{x})$ is *any strictly monotone transformation* of $\eta(\boldsymbol{x})$.

Multi-label classification

1. Show that the Hamming loss regret for multi-label classification can be expressed by the 0/1 regret of individual binary classifiers trained independently for all labels. Furthermore, express the 0/1 regret in terms of the marginal probability.

2. Prove that the ordering of labels according to their decreasing marginal probabilities gives the optimal solution for (unnormalized) rank loss.

3. 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.

4. 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.

5. Derive the time complexity of the ϵ-inference algorithm used for prediction under the subset 0/1 loss in probabilistic classifier chains.