WS 11/12

Introduction to Statistical and Structural Pattern Recognition

Exercise sheet **Bayesian Decision Theory**

Bogdan Savchynskyy - Bogdan.Savchynskyy@iwr.uni-heidelberg.de

All exercises below consider a bayesian problem with given observation x belonging to an observation set $\mathcal{X}$ (feature space),
a finite set $\mathcal{K}$ of object states (latent variables),
a finite set $\mathcal{D}$ of decisions,
a joint probability distribution $p(x,k)$
and a loss function $W\colon \mathcal{K}\times\mathcal{D}\to\mathbb{R}$.

Short exercise formulations specify some of these mathematical objects in more details.

Exercise 1 (**Computing probability of a sum).** Let $\vec{\mathcal{K}} = \mathcal{K}_1\times\mathcal{K}_2\times\cdots\times\mathcal{K}_n$, where $\mathcal{K}_i=\{1,\dots,N\},\ \forall i=1,\dots,n$. Let also $\mathcal{D}=\{d\colon d=\sum_{i=1}^n k_i, k_i\in\mathcal{K}_i\}$.

Let $\vec{\mathcal{K}}(d)=\{\vec{k}\colon\ \sum_{i=1}^n k_i = d\}$. In this case

$$R(d') = \sum_{\vec{k}\in\vec{\mathcal{K}}} p(\vec{k}|x)W(d(\vec{k}),d') = \sum_{d=1}^{\mathcal{D}}\sum_{\vec{k}\in\vec{\mathcal{K}}(d)} p(\vec{k}|x)W(d,d') = \tag{1}$$

$$= \sum_{d=1}^{\mathcal{D}} W(d,d') \underbrace{\sum_{\vec{k}\in\vec{\mathcal{K}}(d)} p(\vec{k}|x)}_{F(d)} = \sum_{d=1}^{\mathcal{D}} F(d)W(d,d'). \tag{2}$$

Construct an algorithm with complexity $O(N^2\cdot n)$ to compute $F(d)$ given $p(\vec{k}|x)$. Write it down in a form of a recursive formula.

Exercise 2 (**Computing probability of a sum).** Let $\vec{\mathcal{K}} = \mathcal{K}_1\times\mathcal{K}_2\times\cdots\times\mathcal{K}_n$, where $\mathcal{K}_i=\{1,\dots,N\},\ \forall i=1,\dots,n$. Let also $\mathcal{D}=\{d\colon d=\sum_{i=1}^n k_i, k_i\in\mathcal{K}_i\}$.

Let $\vec{\mathcal{K}}(d)=\{\vec{k}\colon\ \sum_{i=1}^n k_i = d\}$. In this case

$$R(d') = \sum_{\vec{k}\in\vec{\mathcal{K}}} p(\vec{k}|x)W(d(\vec{k}),d') = \sum_{d=1}^{\mathcal{D}}\sum_{\vec{k}\in\vec{\mathcal{K}}(d)} p(\vec{k}|x)W(d,d') = \tag{3}$$

$$= \sum_{d=1}^{\mathcal{D}} W(d,d') \underbrace{\sum_{\vec{k}\in\vec{\mathcal{K}}(d)} p(\vec{k}|x)}_{F(d)} = \sum_{d=1}^{\mathcal{D}} F(d)W(d,d'). \tag{4}$$

Construct an algorithm with complexity $O(N^2\cdot n)$ to compute $F(d)$ given $p(\vec{k}|x)$. Write it down in a form of a recursive formula.

Remark 0.0.0.1. For an arbitrary loss fuction W an optimal solution

$$d' = \arg\min_{d'=1..|\mathcal{D}|}\sum_{d=1}^{|\mathcal{D}|} F(d)(d-d')^2 \tag{5}$$

can be found in $O(|\mathcal{D}|^2)$ operations by a straight-forward application of (5) given $F(d)$.

Exercise 3 (**Square loss).** Let in conditions of Exercise 2 $W(d,d') = (d-d')^2$.

1. Propose a method(s) for computing (5) in $O(|\mathcal{D}|)$ time.
2. Show, that the solution of (5) corresponds to a math. expectation of the distribution $F(d)$. Use this fact for computing.
3. Computing $F(d)$ according to Exercise 2 requires $O(N^2 \cdot n^2)$ operations. Propose a method for computing (5) in $O(N \cdot n)$ time avoiding computation of $F(d)$. Hint! Use the fact, that math. expectation of a sum is a sum of math. expectations.

Exercise 4 (**Absolute Value Loss).** Let in conditions of Exercise 2 $W(d,d') = |d-d'|$.

1. Propose a method(s) for computing (5) in $O(|\mathcal{D}|)$ time.
2. Show, that the solution of (5) corresponds to a median of the distribution $F(d)$. Use this fact for computing.

Exercise 5 (**Interval loss).** Let in conditions of Exercise 2 $W(d,d') = \left\{ \begin{array}{ll} 0, & |d-d'| \leq \Delta \\ 1, & \text{otherwise} \end{array} \right.$.

1. Propose a method(s) for computing (5) in $O(|D|)$ time (independent of Δ!)
2. Let $\mathcal{D} = \mathcal{D}_x \times \mathcal{D}_y$ contains values in a 2D-grid. Let $W(d,d') = \left\{ \begin{array}{ll} 0, & |d_x - d'_x| + |d_y - d'_y| \leq \Delta \\ 1, & \text{otherwise} \end{array} \right.$.
   Propose a method(s) for computing (5) in $O(|\mathcal{D}|) = O(|\mathcal{D}_x| \cdot |\mathcal{D}_y|)$ time (independent of Δ!)

WS 11/12

Introduction to Statistical and Structural Pattern Recognition

Exercise sheet **Learning Theory**

Bogdan Savchynskyy - Bogdan.Savchynskyy@iwr.uni-heidelberg.de

Exercise 6 (Maximum Likelyhood Estimation of Parameters).
The following four exercises consider a given learning sample $\{(x_i, k_i),\ i = 1 \ldots m\}$. Parameters of given distributions should be found according to a maximum likelihood rule.

1. Gaussian distribution:]

$$p(x|k) = \frac{1}{\sqrt{2\pi}\sigma} \exp - \frac{(x-\mu_k)^2}{2\sigma_k^2}$$

   Find estimation of the mean μ_k and standard deviation σ_k.

2. Conditionally independent features. $\overline{x} = (x_1, \ldots, x_n)$
   $x_i \in X_i,\ i = 1..n$
   $\overline{X} = X_1 \times \cdots \times X_n$
   $\overline{x} \in \overline{X},\ k \in K$
   $p(\overline{x}|k) = \prod_{i=1}^{n} p_i(x_i|k)$
   Find estimation of numbers $p(x|k)$ (non-parametric parameters estimation).

3. Distribution $p(x|k) = \frac{1}{2} e^{-|x-\mu_k|}$ Find estimation of μ_k.

4. Uniform distribution on an interval. $p(x|k) = \begin{cases} \varepsilon, & |x-\mu_k| \le \delta_k \\ 0, & |x-\mu_k| > \delta_k \end{cases}$

   - Find ε, to make $p(x)$ a probability distribution.
   - Find estimations of μ_k and δ_k

Exercise 7 (Straightening of the Feature Space).
The following two exercises address the straightening of the feature space.

1. Separation of sets using a circle. Given 2 finite sets $\mathcal{X}^1$ and $\mathcal{X}^2$ in $\mathbb{R}^2$. It is known that they can be separated by a circle, i.e. $(x_1-\mu_1)^2 + (x_2-\mu_2)^2 > r^2,\ \overline{x} \in X^1$
   $(x_1-\mu_1)^2 + (x_2-\mu_2)^2 < r^2,\ \overline{x} \in X^2$
   However the center μ and the radius r are unknown.

   Construct a straightened space, i.e. space where these points can be separated by a hyperplane. How are coordinates of a normal to the hyperplane in this space connected to μ and r ?
   Consider two situations, when it is known which set should be placed inside the circle and when it is unknown.

2. Consider the target function $f(m_1, m_2, r) = C\frac{m_1 m_2}{r^2}$. Construct a straightened space, i.e. a mapping $x = (m_1, m_2, r) \to \phi(x) = (\phi_1(x), \ldots, \phi_N(x))$ such the corresponding function $F(\phi(x)) = f(x)$ is linear with respect to $\phi(x)$, i.e. can be represented as $F(\phi(x)) = \langle \alpha, \phi(x) \rangle$.

Introduction to Statistical and Structural Pattern Recognition

Exercise sheet **Duality in Convex Programming**

Bogdan Savchynskyy - Bogdan.Savchynskyy@iwr.uni-heidelberg.de

Exercise 8 (Lagrange Dual). Construct the Lagrangian dual for problems:

1. Linearly separable discrimination:

$$\min_{w} \frac{C}{2}\|w\|_2^2 \tag{6}$$
$$\left\langle w, x^i \right\rangle \geq 1,\ i = 1, \dots, m \tag{7}$$

   Show that its dual is also QP and its objective can be represented via a kernel matrix κ, where $\kappa_{ij} = \left\langle x^i, x^j \right\rangle$. Compare this representation to the one cooresponding to a preceptron algorithm.

2. Binary SVM: $\mathcal{X} = \mathcal{X}^0 \cup \mathcal{X}^1$

$$\min_{w,\xi} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} \xi^i \tag{8}$$
$$y_i(\left\langle w, x^i \right\rangle - b) \geq 1 - \xi^i,\ i = 1, \dots, n,\ y_i = 1 \text{ if } x^i \in \mathcal{X}^0 \text{ and } y_i = -1 \text{ if } x^i \in \mathcal{X}^1 \tag{9}$$
$$\xi_i \geq 0 \tag{10}$$

   Show that its dual is also QP and its objective can be represented via a kernel matrix κ, where $\kappa_{ij} = \left\langle x^i, x^j \right\rangle$.

3. Linear programming (LP):

$$\min_{x} \langle \theta, x \rangle \tag{11}$$
$$Ax = b \tag{12}$$

   Show that its dual is also LP.

$$\min_{x} \langle \theta, x \rangle \tag{13}$$
$$Ax = b \tag{14}$$
$$x_1 \geq 0 \tag{15}$$

   Show that its dual is also LP.

$$\min_x \langle \theta, x \rangle \tag{16}$$
$$Ax \geq b \tag{17}$$
$$x \geq 0 \tag{18}$$

Show that its dual is also LP.

$$\min_x x_1 + 2x_3 - x_4 \tag{19}$$
$$x_1 + x_2 = 1 \tag{20}$$
$$x_2 + x_3 + x_4 = 0 \tag{21}$$
$$x_1 \geq 0 \tag{22}$$
$$x_4 \geq 0 \tag{23}$$

Show that its dual is also LP.

Exercise 9 (**Kernel methods (Demo and Discussion)**).
Demo with Gaussian Radial basis functions:
Kernel $\kappa(x, y) = \exp(-\gamma\|x - y\|^2)$
Recognition (inference) $\langle w, \cdot \rangle = \sum_{i=1}^{n} \alpha_i y_i K(x^i, \cdot) + b \gtrless 0$
Learning using dual formulation of the binary SVM (see answers sheet) - just plug a korresponding value of κ_{ij}.
Infinite-dimentional (!) straightening space, where $\kappa(x, y) = \langle \phi(x), \phi(y) \rangle$.

Introduction to Statistical and Structural Pattern Recognition

Exercise sheet **Hidden Markov Chains**

Bogdan Savchynskyy - Bogdan.Savchynskyy@iwr.uni-heidelberg.de

Exercise 10 (**MAP estimation**). The message $\bar{x} = (x_1, x_2, ..., x_n),\ x_i \in A,\ 0 \le i \le n$ represents a sequence of Latin characters and spaces between them. The message is corrupted by a noise in the channel. Each character of the message is distorted independently from others: it stays undistorted with probability $1 - \varepsilon$ and transforms to any other character (which is selected uniformly among others) with probability ε.

Find the most probable undistorted message supposing that the process of creating and distorting of the message can be described by a finite stochastic autonomous automaton.

Required stochastic parameters of the automaton consider to be given.

Exercise 11 (**Locally additive penalty**). MAP estimation of the sequence of hidden variables constitutes a special case of Bayesian theory corresponding to the loss function, which penalizes equally all incorrect inference results. The loss equal to a number of incorrectly recognized characters seems to be more natural in conditions of Exercise 10. Solve this task using such a loss.

Exercise 12 (**Inpainting**). In the communication channel from Exercise 10 some failures happened, which resulted to exchanging of some characters to the special sign "unknown". Find the most probable undistorted message.

Exercise 13 (**Segmentation**). Find the most probable positions of the space character under conditions of Exercise 10.

Exercise 14 (**Median maximization (-0.3 at examination)**). MAP estimation process can be considered as maximizing an average value of $q_i(k_i, k_{i-1}) = \log p_i(x_i, k_i | k_{i-1})$, i.e.

$$\overline{k}^* = \arg\max_{\overline{k} \in \overline{\mathcal{K}}} q_0(k_0) + \sum_{i=1}^{n} q_i(k_i, k_{i-1}) = \arg\max_{\overline{k} \in \overline{\mathcal{K}}} \frac{1}{n+1} (q_0(k_0) + \sum_{i=1}^{n} q_i(k_i, k_{i-1}))$$

Construct an algorithm computing the sequence of hidden variables having a maximal median value.