Prior distribution choices for a mixture of experts *

A. V. Grabovoy[1], **V. V. Strijov**[2]

Abstract: The paper investigates a mixture of expert models. The mixture of experts is a set of experts and the gate function which weighs these experts. Each expert is a linear model. The gate function is a neural network with softmax on the last layer. The paper analyzes different prior distributions for each expert. The authors propose a method that takes into account the relationship between the prior distributions of different experts. The paper uses the EM algorithm for solving the optimization problem. This paper proposes to use the mixture of experts for the problem of circles parameters estimation. Each expert fits one circle in the image. The experiment uses synthetic and real data to test the proposed method. The real data is a human eye image from the iris detection problem.

Keywords: mixture of Experts; bayesian model selection; prior distribution.

1 Introduction

2 Related work

3 Problem statement of circle parameters estimation

This data are binary image

$$\mathbf{M} \in \{0, 1\}^{m_1 \times m_2},$$

where 1 is a black pixel, an image, and 0 is a white pixel, the image background. The image $\mathbf{M}$ is mapped to a set of coordinates $\mathbf{C} = \{x_i, y_i\}_{i=1}^{N}$. The coordinates x_i, y_i is a

*This research was supported by RFBR (project ???) and NTI (project ???).

[1]Moscow Institute of Physics and Technology, grabovoy.av@phystech.edu

[2]Moscow Institute of Physics and Technology, Dorodnicyn Computing Centre, Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences, strijov@phystech.edu

coordinates of black pixels in the image $\mathbf{M}$:

$$\mathbf{C} \in \mathbb{R}^{N \times 2},$$

where N is a number of black pixels in the image $\mathbf{M}$.

Let a pair of coordinates x_0, y_0 is the center of the circle, and r is radius of the circle in the image $\mathbf{M}$.

The points $(x_i, y_i) \in \mathbf{C}$ is a circle locus of points. It satisfies the following condition:

$$(x_i - x_0)^2 + (y_i - y_0)^2 = r^2.$$

Expand brackets:

$$(2x_0) \cdot x_i + (2y_0) \cdot y_i + (r^2 - x_0^2 - y_0^2) \cdot 1 = x_i^2 + y_i^2. \tag{1}$$

Equation (1) is the linear regression problem, given following data:

$$\mathbf{X}\mathbf{w} \approx \mathbf{y}, \quad \mathbf{X} = [\mathbf{C}, \mathbf{1}], \quad \mathbf{y} = [x_1^2 + y_1^2, x_2^2 + y_2^2, \cdots, x_N^2 + y_N^2]^\mathsf{T}, \tag{2}$$

where the parameters $\mathbf{w} = \left[w_1, w_2, w_3\right]^\mathsf{T}$ reconstruct the circle parameters x_0, y_0, r:

$$x_0 = \frac{w_1}{2}, \quad y_0 = \frac{w_2}{2}, \quad r = \sqrt{w_3 + x_0^2 + y_0^2}.$$

The solution of problem (2) reconstructs the circle parameters only if the number of circles in an image is equal to one. If the image consists of several circles, then the authors propose to use a multimodel. The multimodel is an ensemble of linear models. Each linear model approximates only one circle in the image. In this paper, multimodel is a mixture of experts.

4 Problem statement of building a mixture of experts

Generalize one–circle approximation problem to the case of several circles. The data for this case is

$$\mathbf{X} \in \mathbb{R}^{N \times n},$$

where N is a number of datum and n is a number of features.

Definition 4.1. *A multimodel* $\hat{\mathbf{f}}$ *is a mixture of experts if*

$$\hat{\mathbf{f}} = \sum_{k=1}^{K} \pi_k \mathbf{f}_k, \qquad \pi_k(\mathbf{x}, \mathbf{V}) : \mathbb{R}^{n \times |\mathbf{V}|} \to [0, 1], \qquad \sum_{k=1}^{K} \pi_k(\mathbf{x}, \mathbf{V}) = 1, \tag{3}$$

where $\mathbf{f}_k$ *is a local model,* π_k *is a gate function, vector* $\mathbf{w}_k$ *is a parameters of local model and* $\mathbf{V}$ *is a parameters of gate function.*

In this paper, the local model is a linear model. The gate function is a simple 2–layer fully connected neural network.

$$\mathbf{f}_k(\mathbf{x}) = \mathbf{w}_k^{\mathsf{T}}\mathbf{x}, \quad \boldsymbol{\pi}(\mathbf{x}, \mathbf{V}) = \text{softmax}\bigl(\mathbf{V}_1^{\mathsf{T}}\boldsymbol{\sigma}\left(\mathbf{V}_2^{\mathsf{T}}\mathbf{x}\right)\bigr), \tag{4}$$

where $\mathbf{V} = \bigl\{\mathbf{V}_1, \mathbf{V}_2\bigr\}$ is a set of gate function parameters.

Combining (3) and (4), we obtain likelihood:

$$p\bigl(\mathbf{y}, \mathbf{W}|\mathbf{X}, \mathbf{V}\bigr) = \prod_{k=1}^{K} p^k\left(\mathbf{w}_k\right) \prod_{i=1}^{N} \left(\sum_{k=1}^{K} \pi_k p_k\left(y_i|\mathbf{w}_k, \mathbf{x}_i\right)\right), \tag{5}$$

where $\mathbf{W} = \bigl[\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_K\bigr]^{\mathsf{T}}$.

Optimal parameters is a maximum likelihood solution:

$$\hat{\mathbf{W}}, \hat{\mathbf{V}} = \arg\max_{\mathbf{W},\mathbf{V}} p\bigl(\mathbf{y}, \mathbf{W}|\mathbf{X}, \mathbf{V}\bigr). \tag{6}$$

5 EM–algorithm as a solver of optimisation problem

To build a mixture of experts, consider the following probabilistic statement of the problem:

1) a likelihood $p_k\left(y_i|\mathbf{w}_k, \mathbf{x}_i\right) = \mathcal{N}\left(y_i|\mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i, \beta^{-1}\right)$. Parameter β is a level of noise,

2) a prior distribution of parameters $p^k\left(\mathbf{w}_k\right) = \mathcal{N}\left(\mathbf{w}_k|\mathbf{w}_k^0, \mathbf{A}_k\right)$, where $\mathbf{w}_k^0$ is a vector of size $n \times 1$ and $\mathbf{A}_k$ is a covariance matrix,

3) prior regularisation $p\left(\boldsymbol{\varepsilon}_{k,k'}|\boldsymbol{\Xi}\right) = \mathcal{N}\left(\boldsymbol{\varepsilon}_{k,k'}|\mathbf{0}, \boldsymbol{\Xi}\right)$, where $\boldsymbol{\Xi}$ is a covariance matrix and $\boldsymbol{\varepsilon}_{k,k'} = \mathbf{w}_k^0 - \mathbf{w}_{k'}^0$.

Combining assumption $1), 2), 3)$ and equation (5), we obtain

$$\begin{aligned} p\bigl(\mathbf{y}, \mathbf{W}|\mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta\bigr) = &\prod_{k,k'=1}^{K} \mathcal{N}\left(\boldsymbol{\varepsilon}_{k,k'}|\mathbf{0}, \boldsymbol{\Xi}\right) \cdot \\ &\cdot \prod_{k=1}^{K} \mathcal{N}\left(\mathbf{w}_k|\mathbf{w}_k^0, \mathbf{A}_k\right) \prod_{i=1}^{N} \left(\sum_{k=1}^{K} \pi_k \mathcal{N}\left(y_i|\mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i, \beta^{-1}\right)\right), \end{aligned} \tag{7}$$

where $\mathbf{A} = \bigl\{\mathbf{A}_1, \cdots, \mathbf{A}_K\bigr\}$.

Let us introduce a binary matrix $\mathbf{Z}$. Element z_{ik} is equal to 1 if and only if object i related to the local model k. If we combine the binary matrix $\mathbf{Z}$ with logarithm of likelihood (7), we get

$$
\begin{aligned}
\log p\big(\mathbf{y}, \mathbf{Z}, \mathbf{W} | \mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta\big) = & \\
= \sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} \left[\log \pi_k (\mathbf{x}_i, \mathbf{V}) - \frac{\beta}{2} \left(y_i - \mathbf{w}_k^{\mathsf{T}} \mathbf{x}_i\right)^2 + \frac{1}{2} \log \frac{\beta}{2\pi}\right] + & \\
+ \sum_{k=1}^{K} \left[-\frac{1}{2} \left(\mathbf{w}_k - \mathbf{w}_k^0\right)^{\mathsf{T}} \mathbf{A}_k^{-1} \left(\mathbf{w}_k - \mathbf{w}_k^0\right) + \frac{1}{2} \log \det \mathbf{A}_k^{-1} - \frac{n}{2} \log 2\pi\right] + & \\
+ \sum_{k=1}^{K} \sum_{k'=1}^{K} \left[-\frac{1}{2} \left(\mathbf{w}_k^0 - \mathbf{w}_{k'}^0\right)^{\mathsf{T}} \boldsymbol{\Xi}^{-1} \left(\mathbf{w}_k^0 - \mathbf{w}_{k'}^0\right) + \frac{1}{2} \log \det \boldsymbol{\Xi} - \frac{n}{2} \log 2\pi\right]. &
\end{aligned} \tag{8}
$$

Combining (6) and (8), we obtain new optimisation problem:

$$
\mathbf{W}, \mathbf{Z}, \mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta = \arg \max_{\mathbf{W}, \mathbf{Z}, \mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta} \log p\big(\mathbf{y}, \mathbf{Z}, \mathbf{W} | \mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta\big). \tag{9}
$$

In this paper, we are using expectation–maximization [23] algorithm for finding maximum likelihood solution for the multimodel with latent variables $\mathbf{Z}$.

E–step. Let a join distribution $q(\mathbf{Z}, \mathbf{W})$ satisfy the assumption of mean field approximation $q(\mathbf{Z}, \mathbf{W}) = q(\mathbf{Z})\, q(\mathbf{W})$ [23]. In the text below, the symbol $\propto$ means that both sides are equal to up to an additive constant. Firstly, find the distribution $q(\mathbf{Z})$:

$$
\begin{aligned}
\log q(\mathbf{Z}) &= \mathsf{E}_{q/\mathbf{Z}} \log p\big(\mathbf{y}, \mathbf{Z}, \mathbf{W} | \mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta\big) \propto \\
&\propto \sum_{i+1}^{N} \sum_{k=1}^{K} z_{ik} \left[\log \pi_k (\mathbf{x}_i, \mathbf{V}) - \frac{\beta}{2} \left(y_i^2 - \mathbf{x}_i^{\mathsf{T}} \mathsf{E}\mathbf{w}_k + \mathbf{x}_i^{\mathsf{T}} \mathsf{E}\mathbf{w}_k \mathbf{w}_k^{\mathsf{T}} \mathbf{x}_i\right) + \frac{1}{2} \log \frac{\beta}{2\pi}\right] \\
p(z_{ik} = 1) &= \frac{\exp\left(\log \pi_k (\mathbf{x}_i, \mathbf{V}) - \frac{\beta}{2} \left(\mathbf{x}_i^{\mathsf{T}} \mathsf{E}\mathbf{w}_k \mathbf{w}_k^{\mathsf{T}} \mathbf{x}_i - \mathbf{x}_i^{\mathsf{T}} \mathsf{E}\mathbf{w}_k\right)\right)}{\sum_{k'=1}^{K} \exp\left(\log \pi_{k'} (\mathbf{x}_i, \mathbf{V}) - \frac{\beta}{2} \left(\mathbf{x}_i^{\mathsf{T}} \mathsf{E}\mathbf{w}_{k'} \mathbf{w}_{k'}^{\mathsf{T}} \mathbf{x}_i - \mathbf{x}_i^{\mathsf{T}} \mathsf{E}\mathbf{w}_{k'}\right)\right)}.
\end{aligned} \tag{10}
$$

The distribution $q\big(z_{ik}\big)$ is a Bernoulli distribution with the probability z_{ik} from equation (10). Secondly, find the distribution $q(\mathbf{W})$:

$$
\begin{aligned}
\log q(\mathbf{W}) &= \mathsf{E}_{q/\mathbf{W}} \log p\big(\mathbf{y}, \mathbf{Z}, \mathbf{W} | \mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta\big) \propto \\
&\propto \sum_{i=1}^{N} \sum_{k=1}^{K} \mathsf{E} z_{ik} \left[\log \pi_k (\mathbf{x}_{i,\mathbf{V}}) - \frac{\beta}{2} \left(y_i - \mathbf{w}_k^{\mathsf{T}} \mathbf{x}_i\right)^2 + \frac{1}{2} \log \frac{\beta}{2\pi}\right] + \\
&+ \sum_{k=1}^{K} \left[-\frac{1}{2} \left(\mathbf{w}_k - \mathbf{w}_k^0\right)^{\mathsf{T}} \mathbf{A}_k^{-1} \left(\mathbf{w}_k - \mathbf{w}_k^0\right) + \frac{1}{2} \log \det \mathbf{A}_k^{-1} - \frac{n}{2} \log 2\pi\right] \\
&\propto \sum_{k=1}^{K} \left[\mathbf{w}_k^{\mathsf{T}} \left(\mathbf{A}_k^{-1} \mathbf{w}_k^0 + \beta \sum_{i=1}^{N} \mathbf{x}_i y_i \mathsf{E} z_{ik}\right) - \frac{1}{2} \mathbf{w}_k^{\mathsf{T}} \left(\mathbf{A}_k^{-1} + \beta \sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^{\mathsf{T}}\right) \mathbf{w}_k\right].
\end{aligned}
$$

The distribution $q(\mathbf{w}_k)$ is a normal distribution with a mean $\mathbf{m}_k$ and a covariance matrix $\mathbf{B}_k$. The distribution parameters $\mathbf{m}_k, \mathbf{B}_k$ are therefore given by

$$\mathbf{m}_k = \mathbf{B}_k \left(\mathbf{A}_k^{-1}\mathbf{w}_k^0 + \beta \sum_{i=1}^{N} \mathbf{x}_i y_i \mathsf{E} z_{ik} \right), \qquad \mathbf{B}_k = \left(\mathbf{A}_k^{-1} + \beta \sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^{\mathsf{T}} \mathsf{E} z_{ik} \right)^{-1}.$$

M–step. In the M step, the distribution $q(\mathbf{Z}, \mathbf{W})$ is held fixed and the lower bound $\mathcal{L}\left(\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta\right)$ is maximized with respect to parameters $\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta$:

$$\begin{aligned}
\mathcal{L}\left(\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta\right) &= \mathsf{E}_q \log p\big(\mathbf{y}, \mathbf{Z}, \mathbf{W} | \mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta\big) = \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} \mathsf{E} z_{ik} \left[\log \pi_k(\mathbf{x}_i, \mathbf{V}) - \frac{\beta}{2} \mathsf{E}\left(y_i - \mathbf{w}_k^{\mathsf{T}} \mathbf{x}_i\right)^2 + \frac{1}{2}\log\frac{\beta}{2\pi} \right] + \\
&+ \sum_{k=1}^{K} \left[-\frac{1}{2} \mathsf{E}\left(\mathbf{w}_k - \mathbf{w}_k^0\right)^{\mathsf{T}} \mathbf{A}_k^{-1} \left(\mathbf{w}_k - \mathbf{w}_k^0\right) + \frac{1}{2}\log\det \mathbf{A}_k^{-1} - \frac{n}{2}\log 2\pi \right] + \\
&+ \sum_{k=1}^{K}\sum_{k'=1}^{K} \left[-\frac{1}{2} \left(\mathbf{w}_k^0 - \mathbf{w}_{k'}^0\right)^{\mathsf{T}} \boldsymbol{\Xi}^{-1} \left(\mathbf{w}_k^0 - \mathbf{w}_{k'}^0\right) + \frac{1}{2}\log\det\boldsymbol{\Xi} - \frac{n}{2}\log 2\pi \right].
\end{aligned} \tag{11}$$

To find the optimal parameters $\mathbf{V}$ we use the gradient optimization method. The gradient method guarantees convergence to local extrema. Using (11), we get

$$\begin{aligned}
\frac{\partial \mathcal{L}\left(\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta\right)}{\partial \mathbf{A}_k^{-1}} &= \frac{1}{2}\mathbf{A}_k - \frac{1}{2}\mathsf{E}\left(\mathbf{w}_k - \mathbf{w}_k^0\right)\left(\mathbf{w}_k - \mathbf{w}_k^0\right)^{\mathsf{T}} = 0, \\
\mathbf{A}_k &= \mathsf{E}\mathbf{w}_k\mathbf{w}_k^{\mathsf{T}} - \mathbf{w}_k^0 \mathsf{E}\mathbf{w}_k^{\mathsf{T}} - \mathsf{E}\mathbf{w}_k \mathbf{w}_k^{0\mathsf{T}} + \mathbf{w}_k^0 \mathbf{w}_k^{0\mathsf{T}}.
\end{aligned}$$

Similarly, we get

$$\begin{aligned}
\frac{\partial \mathcal{L}\left(\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta\right)}{\partial \beta} &= \sum_{k=1}^{K}\sum_{i=1}^{N} \left(\frac{1}{\beta}\mathsf{E} z_{ik} - \frac{1}{2}\mathsf{E} z_{ik} \left[y_i^2 - 2y_i \mathbf{x}_i^{\mathsf{T}} \mathsf{E}\mathbf{w}_k + \mathbf{x}_i^{\mathsf{T}} \mathbf{w}_k \mathbf{w}_k^{\mathsf{T}} \mathbf{x}_i \right] \right) = 0, \\
\frac{1}{\beta} &= \frac{1}{N} \sum_{i=1}^{N}\sum_{k=1}^{K} \left[y_i^2 - 2y_i \mathbf{x}_i^{\mathsf{T}} \mathsf{E}\mathbf{w}_k + \mathbf{x}_i^{\mathsf{T}} \mathsf{E}\mathbf{w}_k \mathbf{w}_k^{\mathsf{T}} \mathbf{x}_i \right] \mathsf{E} z_{ik}.
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \mathcal{L}\left(\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta\right)}{\partial \mathbf{w}_k^0} &= \mathbf{A}_k^{-1}\big(\mathsf{E}\mathbf{w}_k - \mathbf{w}_k^0\big) + \boldsymbol{\Xi} \sum_{k'=1}^{K} \big[\mathbf{w}_{k'}^0 - \mathbf{w}_k^0\big] = 0, \\
\mathbf{w}_k^0 &= \left[\mathbf{A}_k^{-1} + (K-1)\,\boldsymbol{\Xi}\right]^{-1} \left(\mathbf{A}_k^{-1}\mathsf{E}\mathbf{w}_k + \boldsymbol{\Xi} \sum_{k'=1,\ k'\neq k}^{K} \mathbf{w}_{k'}^0 \right).
\end{aligned} \tag{12}$$

The formulas (10–12) are an iterative procedure which convergence to a local maximum of optimisation problem (9). If in the list of probabilistic statements we leave only the statement 1) then we find a solution to the optimization problem (7). If in the list of probabilistic statements we leave statements 1) and 2) then we find a solution to the optimization problem with prior distribution on each local model parameters. If in the list of probabilistic statements we leave all statements 1), 2) and 3) then we find a solution to the optimization problem (9) with prior distributions and relationships between prior distributions of different local models.

6 Computational experiment

7 Conclusion

References

[1] *Chen Tianqi, Guestrin Carlos* XGBoost: A Scalable Tree Boosting System // KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016.

[2] *Chen Xi, Ishwaran Hemant* Random Forests for Genomic Data Analysis // Genomics. 2012. Issues. 99, No 6. pp. 323–329.

[3] *Yuksel Seniha Esen, Wilson Joseph N., Gader Paul D* Twenty Years of Mixture of Experts // IEEE Transactions on Neural Networks and Learning Systems. 2012. Issues. 23, No 8. pp. 1177–1193.

[4] *Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz* Outrageously large neural networks: the sparsely-gated mixture-of-experts layer // ICLR, 2017.

[5] *Rasmussen Carl Edward, Ghahramani Zoubin* Infinite Mixtures of Gaussian Process Experts // Advances in Neural Information Processing Systems 14. 2002. pp. 881–888.

[6] *M. I. Jordan* Hierarchical mixtures of experts and the EM algorithm // Neural Comput., vol. 6, no. 2, pp. 181–214, 1994.

[7] *C. A. M. Lima, A. L. V. Coelho, F. J. Von Zuben* Hybridizing mixtures of experts with support vector machines: Investigation into nonlinear dynamic systems identification // Inf. Sci., vol. 177, no. 10, pp. 2049–2074, 2007.

[8] *L. Cao* Support vector machines experts for time series forecasting // Neurocomputing, vol. 51, pp. 321–339, Apr. 2003.

[9] *A. P. Dempster, N. M. Laird and D. B. Rubin* Maximum Likelihood from Incomplete Data via the EM Algorithm // Journal of the Royal Statistical Society. Series B (Methodological), Vol. 39, No. 1 pp. 1-38, 1977.

[10] *M. I. Jordan, R. A. Jacobs* Hierarchies of adaptive experts // in Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 1991, pp. 985–992.

[11] *M. S. Yumlu, F. S. Gurgen, N. Okay* Financial time series prediction using mixture of experts // in Proc. 18th Int. Symp. Comput. Inf. Sci., 2003, pp. 553–560.

[12] *Y. M. Cheung, W. M. Leung, and L. Xu* Application of mixture of experts model to financial time series forecasting // in Proc. Int. Conf. Neural Netw. Signal Process., 1995, pp. 1–4.

[13] *A. S. Weigend, S. Shi* Predicting daily probability distributions of S&P500 returns // J. Forecast., vol. 19, no. 4, pp. 375–392, 2000.

[14] *R. Ebrahimpour, M. R. Moradian, A. Esmkhani, F. M. Jafarlou* Recognition of Persian handwritten digits using characterization loci and mixture of experts // J. Digital Content Technol. Appl., vol. 3, no. 3, pp. 42–46, 2009.

[15] *A. Estabrooks, N. Japkowicz* A mixture-of-experts framework for text classification //in Proc. Workshop Comput. Natural Lang. Learn., Assoc. Comput. Linguist., 2001, pp. 1–8.

[16] *S. Mossavat, O. Amft, B. de Vries, P. Petkov, W. Kleijn* A Bayesian hierarchical mixture of experts approach to estimate speech quality // in Proc. 2nd Int. Workshop Qual. Multimedia Exper., pp. 200–205., 2010

[17] *F. Peng, R. A. Jacobs, M. A. Tanner* Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition // J. Amer. Stat. Assoc., vol. 91, no. 435, pp. 953–960, 1996.

[18] *A. Tuerk* The state based mixture of experts HMM with applications to the recognition of spontaneous speech // Ph.D. thesis, Dept. Eng., Univ. Cambridge, Cambridge, U.K., 2001.

[19] *C. Sminchisescu, A. Kanaujia, and D. Metaxas* B M3 E: Discrimina- tive density propagation for visual tracking // IEEE Trans. Pattern Anal. Mach. Intell., vol. 29, no. 11, pp. 2030–2044, 2007.

[20] *I. Matveev* Detection of iris in image by interrelated maxima of brightness gradient projections // Appl.Comput. Math. 9 (2), 252–257, 2010.

[21] *I. Matveev, I. Simonenko.* Detecting precise iris boundaries by circular shortest path method // Pattern Recognition and Image Analysis. 24. 304-309. 2014.

[22] *K. Bowyer, K. Hollingsworth, P. Flynn* A Survey of Iris Biometrics Research: 2008–2010.

[23] *Bishop C.* Pattern Recognition and Machine Learning. — Berlin: Springer, 2006. 758 p.