Priori distribution choices for a mixture of experts *

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

Abstract: The paper investigates a mixture of expert models. A mixture of expert is a set of experts and gate function which weighs these experts. Each expert is a linear model. A gate function is a neural network with softmax on the last layer. In this article, we analysed different a priori distributions for each expert. We proposed a method that consider the relationship between the a priori distributions of different experts. To solve the obtained optimization problem, we used the EM algorithm. The problem of finding circles in the image is considered as the task of a mixture of experts.Each circle in the image is a one expert. Our method was tested on the real and synthetic data. A real data is a human eyes from rhe iris detection problem.

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

1 Introduction

2 Related work

3 Problem statement of finding the circle parameters in the image

Binary image are considered as data:

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

*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

where 1 is a black point — an image and 0 is a white point — image background. The set $\mathbf{C}$ constructed by mapping the image $\mathbf{M}$ to the set of coordinates x_i, y_i of black points in the image:

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

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

Let x_0, y_0 is a coordinates of center and r is a radius of circle in the image $\mathbf{M}$. Circle parameters x_0, y_0, r need to be found.

The points $(x_i, y_i) \in \mathbf{C}$ is geometric locus of circle points. This locus approximated by using circle equation:

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

Rewrite brackets:

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

The equation (1) is a linear regression problem with 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 parameters $\mathbf{w} = \left[w_1, w_2, w_3\right]^{\mathsf{T}}$ reconstruct 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) reconstruct circle parameters only when number of circles in image is one. If the image has several circles, then it is proposed to use a multimodel. The multimodel consists of linear models. Each linear model approximates only one circle in the image. As a multimodel, we consider a mixture of experts model.

4 Problem statement of building a mixture of experts

Let generalize one circle approximation problem in the case when there are several circles. Consider the data from the equation (2) without assumption that circle is one:

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

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

Definition 4.1. *A mixture of experts is a multimodel in which the weight of each local model depends on the object:*

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

where $\hat{\mathbf{f}}$ *— multimodel,* $\mathbf{f}_k$ *— local models,* π_k *— gate function,* $\mathbf{w}_k$ *— parameters of local model and* $\mathbf{V}$ *— parameters of gate function.*

The linear models are considered as local models. A simple 2–layer fully connected neural network is considered as a gate function:

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

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

All parameters optimises according to the maximum likelihood principle:

$$p\bigl(\mathbf{y}, \mathbf{W}|\mathbf{X}, \mathbf{V}\bigr) = \prod_{k=1}^{K}p^k\left(\mathbf{w}_k\right)\prod_{i=1}^{N}\left(\sum_{k=1}^{K}\pi_k p_k\left(y_i|\mathbf{w}_k, \mathbf{x}_i\right)\right), \tag{3}$$

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

The optimisation problem for finding optimal parameters of local models and optimal mixture parameters:

$$\hat{\mathbf{W}}, \hat{\mathbf{V}} = \arg\max_{\mathbf{W}, \mathbf{V}} p\bigl(\mathbf{y}, \mathbf{W}|\mathbf{X}, \mathbf{V}\bigr). \tag{4}$$

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

2) a priori 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 $\mathbf{w}_k^0$ — a vector of size $n \times 1$, $\mathbf{A}_k$ — covariance matrix of parameters,

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

Under the previous assumption, the likelihood (3) is rewritten to:

$$\begin{aligned} p\bigl(\mathbf{y}, \mathbf{W}|\mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta\bigr) = &\prod_{k,k'=1}^{K}\mathcal{N}\left(\boldsymbol{\varepsilon}_{k,k'}|\mathbf{0}, \boldsymbol{\Xi}\right)\cdot \\ &\cdot\prod_{k=1}^{K}\mathcal{N}\left(\mathbf{w}_k|\mathbf{w}_k^0, \mathbf{A}_k\right)\prod_{i=1}^{N}\left(\sum_{k=1}^{K}\pi_k\mathcal{N}\left(y_i|\mathbf{w}_k^\mathsf{T}\mathbf{x}_i, \beta^{-1}\right)\right), \end{aligned} \tag{5}$$

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

Consider the matrix of hidden variables $\mathbf{Z}$ for solving problem (4) under assumption (5). In matrix $\mathbf{Z}$ all $z_{ik} = 1$ if and only if object i related to local model k and $z_{ik} = 0$ otherwise.

Logarithm of likelihood (5) rewrites to following view by using matrix $\mathbf{Z}$:

$$\begin{aligned}
&\log p\bigl(\mathbf{y}, \mathbf{Z}, \mathbf{W}|\mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta\bigr) = \\
&\qquad = \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik}\left[\log \pi_k\left(\mathbf{x}_i, \mathbf{V}\right) - \frac{\beta}{2}\left(y_i - \mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i\right)^2 + \frac{1}{2}\log\frac{\beta}{2\pi}\right] + \\
&\qquad + \sum_{k=1}^{K}\left[-\frac{1}{2}\left(\mathbf{w}_k - \mathbf{w}_k^0\right)^{\mathsf{T}}\mathbf{A}_k^{-1}\left(\mathbf{w}_k - \mathbf{w}_k^0\right) + \frac{1}{2}\log\det\mathbf{A}_k^{-1} - \frac{n}{2}\log 2\pi\right] + \\
&\qquad + \sum_{k=1}^{K}\sum_{k'=1}^{K}\left[-\frac{1}{2}\left(\mathbf{w}_k^0 - \mathbf{w}_{k'}^0\right)^{\mathsf{T}}\boldsymbol{\Xi}^{-1}\left(\mathbf{w}_k^0 - \mathbf{w}_{k'}^0\right) + \frac{1}{2}\log\det\boldsymbol{\Xi} - \frac{n}{2}\log 2\pi\right].
\end{aligned} \tag{6}$$

The optimisation problem (4) for log–likelihood (6) rewrites as follows

$$\mathbf{W}, \mathbf{Z}, \mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta = \arg\max_{\mathbf{W},\mathbf{Z},\mathbf{V},\mathbf{W}^0,\mathbf{A},\beta} \log p\bigl(\mathbf{y}, \mathbf{Z}, \mathbf{W}|\mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta\bigr). \tag{7}$$

To find a local minimum in the optimization problem (7), we use the variational EM–algorithm [23].

E–step. Let the variational distribution $q\left(\mathbf{Z}, \mathbf{W}\right)$ satisfy the assumption [23] of mean field approximation $q\left(\mathbf{Z}, \mathbf{W}\right) = q\left(\mathbf{Z}\right)q\left(\mathbf{W}\right)$. Find the variational distribution $q\left(\mathbf{Z}, \mathbf{W}\right)$ closest to $p\bigl(\mathbf{Z}, \mathbf{W}|\mathbf{y}, \mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta\bigr)$. In the text below, the symbol $\propto$ means equality up to an additive constant. First, find the distribution of the hidden variable $q\left(\mathbf{Z}\right)$:

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

The distribution $q\bigl(z_{ik}\bigr)$ is a Bernoulli distribution with probability z_{ik} from the equation (8). Second, find the distribution of parameters $q\left(\mathbf{W}\right)$

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

The distribution $q(\mathbf{w}_k)$ is a normal distribution with mean $\mathbf{m}_k$ and covariance matrix $\mathbf{B}_k$. The distribution parameters $\mathbf{m}_k, \mathbf{B}_k$ are calculated as follows:

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

M–step. Find hyperparameters $\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta$ from maximisation expected value of log-likelihood under the condition of a variational distribution $q(\mathbf{Z}, \mathbf{W})$.

$$\begin{aligned}\mathcal{F}\left(\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta\right) &= \mathsf{E}_q \log p\big(\mathbf{y}, \mathbf{Z}, \mathbf{W} | \mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \Xi, \beta\big) = \\ &= \sum_{i=1}^{N}\sum_{k=1}^{K} \mathsf{E} z_{ik} \left[\log \pi_k(\mathbf{x}_i, \mathbf{V}) - \frac{\beta}{2} \mathsf{E}\left(y_i - \mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i\right)^2 + \frac{1}{2}\log\frac{\beta}{2\pi} \right] + \\ &+ \sum_{k=1}^{K} \left[-\frac{1}{2}\mathsf{E}\left(\mathbf{w}_k - \mathbf{w}_k^0\right)^{\mathsf{T}} \mathbf{A}_k^{-1}\left(\mathbf{w}_k - \mathbf{w}_k^0\right) + \frac{1}{2}\log\det \mathbf{A}_k^{-1} - \frac{n}{2}\log 2\pi \right] + \\ &+ \sum_{k=1}^{K}\sum_{k'=1}^{K} \left[-\frac{1}{2}\left(\mathbf{w}_k^0 - \mathbf{w}_{k'}^0\right)^{\mathsf{T}} \boldsymbol{\Xi}^{-1}\left(\mathbf{w}_k^0 - \mathbf{w}_{k'}^0\right) + \frac{1}{2}\log\det\boldsymbol{\Xi} - \frac{n}{2}\log 2\pi \right].\end{aligned} \tag{9}$$

To find the parameters $\mathbf{V}$, which maximising the function (9), we use the gradient optimization method. This method guarantees convergence to local extrema. Let find the parameters $\mathbf{A}_k$, which maximising the function (9):

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

Similarly, we find optimal value of β and $\mathbf{w}_0^k$.

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

$$\begin{aligned}\frac{\partial \mathcal{F}\left(\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta\right)}{\partial \mathbf{w}_k^0} &= \mathbf{A}_k^{-1}\big(\mathsf{E}\mathbf{w}_k - \mathbf{w}_k^0\big) + \boldsymbol{\Xi}\sum_{k'=1}^{K}\left[\mathbf{w}_{k'}^0 - \mathbf{w}_k^0\right] = 0, \\ \mathbf{w}_k^0 &= \left[\mathbf{A}_k^{-1} + (K-1)\boldsymbol{\Xi}\right]^{-1}\left(\mathbf{A}_k^{-1}\mathsf{E}\mathbf{w}_k + \boldsymbol{\Xi}\sum_{k'=1,\ k'\neq k}^{K}\mathbf{w}_{k'}^0\right).\end{aligned} \tag{10}$$

The formulas (8–10) are an iterative procedure which convergence to local maximum of optimisation problem (7). If in the list of probabilistic statements we consider only the statement 1) then find solution the optimisation problem (5) without any priori distribution. If in the list of probabilistic statements we considers statements 1) and 2) then find solution the optimisation problem with a priori distribution on the local models parameters. If in the list of probabilistic statements we considers all statements 1), 2) and 3) then find solution the optimisation problem (7) with a priori distributions and relationships between a priori 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.