COSC 7362 Advanced Machine Learning (*Dr. Eick*)
Quiz1
Solution Sketches
Mo., October 26, 2015

Your Name:
Your Student id:

Problem 1 --- Anomaly Detection (& Mixtures of Gaussians) [22]
Problem 2 --- Density Estimation [20]
Problem 3 --- Backpropagation Algorithm [8]
Problem 4 --- Deep Learning [17]

Σ [67]:

Grade:

The exam is "open books" and you have 70 minutes to complete the exam.
The exam will count approx. 18-23% towards the course grade.

1) Anomaly Detection & Mixture of Gaussians (concerns Irad Ben-Gal and Punzo et al. papers)

a) Give a definition of the term 'outlier'—*just one is sufficient*! [2]

No answer given!

b) Give a sketch of a distance-based outlier detection method, of your own preference. [4]

No answer given!

c) What advantages do you see in using model-based outlier detection methods over its competitors? [3]

- **allows for the probabilistic assessment of objects being outliers (allows to rank observations based on their likelihood being outliers)**
- **sound theoretical foundation**
- **fast online performance (model is obtained offline)**
- **deriving model parameters is well automated (does not depend on choosing other meta parameters**

d) Why is it desirable to have robust outlier detection techniques? [2]

A robust method is not much affected by a small number of extreme or erroneous observations

e) What is swamping? Give an example of a swamping effect! [4]

No answer given!

f) Gaussian mixture models are much more popular than ordinary Gaussians in model-based outlier detection. Why do you believe is this the case? [3]

Ordinary Gaussians assume unimodal and symmetric density functions; the distribution of most datasets frequently violate those assumptions; mixtures are more flexible in the types of density functions they support by allowing for multi-modal, non-symmetric density functions.

g) Give a high-level verbal description of the outlier detection approach that is described in the Punzo et al. paper. Limit your description to 3-6 sentences. [4]

Consider a mixture gaussians, where each component is:
2 gaussian with the same mean, but different variance and different weights (α, and $(1-\alpha)$). different α for different component. For an observation:
- find the component i that best fit
- In that component i, find the best fit gaussian $\alpha_i g_i$ or $(1-\alpha_i) g_i$

If the fit is a "bad" one: large variance → outlier

2) Density Estimation (covering the Silverman and the Vandermeulen et al. paper)

a) Assume you use the naïve estimator, described in Section 2.3 of the Silverman paper, for the dataset {1.3, 1.4, 2.1, 2.4, 3.6, 3.8}; moreover, assume h=0.5; compute the density in 1.7 and 4.0; that is, compute f(1.7) and f(4.0)! [4]

$n=6$ $\hat{f}(x) = \frac{1}{6}\sum_{i=1}^{6} w\left(\frac{x - X_i}{0.5}\right)\cdot\frac{1}{0.5} = \frac{1}{3}\sum_{i=1}^{6} w\left((x - X_i).2\right)$

$\hat{f}(1.7) = \frac{1}{3}\left[1/2 + 1/2 + 1/2\right] = \frac{1}{2}$ $\hat{f}(4.0) = \frac{1}{3}\left[\frac{1}{2} + \frac{1}{2}\right] = \frac{1}{3}$

b) What advantage do you see in using the kernel density estimator over the naïve estimator? [2]
obtains a smoother, differentiable and continuous density function

c) Both estimators use a parameter h; how does the selection of h impact the shape of the obtained density function? [3]
h large: obtain smooth density functions with very few peaks and less variation→captures the global/regional characteristics of a dataset
h large: obtain spiky, rough density function with more variation and a lot of maxima →capture the local characteristics of a dataset

d) What is the idea of the *maximum penalized likelihood estimator* and what advantages do you see in using this approach? [4]

The idea is to use maximum likelihood to assess the goodness of the model fit and to penalize model complexity (measured by the roughness of the density functions), by formulating obtaining the "best" density function as a multi-objective optimization problem that balances the two objectives.

e) What is the high-level goal of the research that is described in Vandermeulen et al.'s paper? Limit you answer to 2-4 sentences! [3]

Building a robust kernel estimator: robust in the presence of noise. The assumption is limited to a special form of noise: bounded and "nearly" uniform distributed over the dataset. The idea is to eliminate noise by scalling and projection.

f) Give a verbal description (4-6 sentences) of the contamination procedure that is proposed in the Vandermeulen et al. paper [4]

see page 3 of the paper

3) Backpropagation [8]

a) Why is weight-updating of multi-layered neural networks more complicated than updating weights of a perceptron? [3]

The error of the nodes of the intermediate layers is not known in advance and has to be estimated.

b) When using the back propagation algorithm—what impacts the degree of weight update; which connections receive larger weight updates? [3]

...learning rate, input activation, associated error

c) When does a connection receive a weight update of 0—that is, its weight does not change? [2]

no error

4) Deep Learning [17]

a) What are the unique characteristics of deep learning approaches, if compared to other approaches in machine learning? Limit your answer to 4-6 sentences! [4]

No answer given!

b) What is an auto-encoder? [3]

Framework parametrized through encoder (allows straightforward computation of feature vector h from input and decoder (maps feature space back to input space. Parameters for both are learnt at the same time and try to find which result in the lower reconstruction errors

c) What is the purpose of developing auto-encoders? What role do they play in machine learning? [4]

to discover hidden (latent) variables.

To overcome the curse of dimensionality, it is widely used in pattern recognition related to images, audio, natural languages. Learning a function of 200x200 inputs (for an image) is untractable, auto encoders help to reduce the dimension only consider a low number of hidden features.

4 good

d) Bengio et al. suggest incorporating *Generic AI-specific Priors* into Deep Learning architectures. Why do Bengio et al. believe this is important? Give an example of an AI-specific prior whose incorporation into a Deep Learning architecture would be beneficiary! [6]

No answer given, **but it was rarely mentioned in the answers that using AI-specific priors frequently lead to a simplification of algorithms and architectures or in a reduction of parameters that need to be estimated when employed in deep learning systems.**