Exercises
Introduction to Machine Learning
Spring 2022

Institute for Machine learning
Dept. of Computer Science, ETH Zürich
Prof. Dr. Andreas Krause, Prof. Dr. Fanny Yang

Homework 6
(Generative models, GMM's and GAN's)

For questions, please refer to Moodle.
Released on **23 May, 2022**

General Instructions

- Submission of solutions is not mandatory but solving the exercises are highly recommended. The master solution will be released next week.
- Part of the exercises are available on Moodle. These problems are marked with [✓].

Exercise 1: Discriminative and Generative Models

(a) [✓] Which of these models are generative?

1. Log. Regression 2. SVM 3. Neural Networks 4. Naive Bayes classifiers

Suppose you want to estimate labels Y based on features X. Which of the following probability distributions can you estimate if you use ...

(b) [✓] ... a discriminative model?

1. $P(X,Y)$ 2. $P(Y|X)$ 3. $P(X|Y)$ 4. $P(X)$ 5. $P(Y)$

(c) [✓] ... a generative model?

1. $P(X,Y)$ 2. $P(Y|X)$ 3. $P(X|Y)$ 4. $P(X)$ 5. $P(Y)$

Now suppose you decide to use a generative model. For this you explicitly model the prior $P(Y)$ and the likelihood $P(X|Y)$.

(d) How do you calculate all other probability distributions that you can estimate with this model?

Suppose you use a Gaussian Bayes Classifier for binary classification ($y \in \{-1, +1\}$).

A Gaussian Bayes Classifier explicitly models the prior $P(Y = y)$ and likelihood $p_{X|Y}(x|y)$. The prior is modeled with a categorical distribution and the likelihood is modeled according to

$$p_{X|Y}(x|y) = \mathcal{N}(x; \mu_y, \Sigma_y).$$

Which assumptions have to hold so that your model can also be named ...

(e) [✓] ... a Linear Discriminant Analysis (LDA)?

1. $P(Y = y) = 1/2$ 2. $\Sigma_y = \sigma_y^2 \cdot \mathbb{I}$ 3. $\Sigma_y = \text{diag}(\sigma_{y,1}^2, \sigma_{y,2}^2, \cdots)$ 4. $\Sigma_{+1} = \Sigma_{-1}$ 5. none

(f) [✓] ... a Fisher's Linear Discriminant Analysis?

1. $P(Y = y) = 1/2$ 2. $\Sigma_y = \sigma_y^2 \cdot \mathbb{I}$ 3. $\Sigma_y = \text{diag}(\sigma_{y,1}^2, \sigma_{y,2}^2, \cdots)$ 4. $\Sigma_{+1} = \Sigma_{-1}$ 5. none

(g) [✓] ... a Quadratic Discriminant Analysis (QDA)?

1. $P(Y = y) = 1/2$ 2. $\Sigma_y = \sigma_y^2 \cdot \mathbb{I}$ 3. $\Sigma_y = \text{diag}(\sigma_{y,1}^2, \sigma_{y,2}^2, \cdots)$ 4. $\Sigma_{+1} = \Sigma_{-1}$ 5. none

(h) [ ] . . . a Gaussian Naive Bayes Classifier?

1. $P(Y = y) = 1/2$ 2. $\Sigma_y = \sigma_y^2 \cdot \mathbb{I}$ 3. $\Sigma_y = \text{diag}(\sigma_{y,1}^2, \sigma_{y,2}^2, \cdots)$ 4. $\Sigma_{+1} = \Sigma_{-1}$ 5. none

(i) [ ] With generative modelling one can explicitly include a bias in the model by defining the structure of the likelihood $P(X|Y)$. ☐ True ☐ False

Suppose you got a very large data set $\{(x_i, y_i)\}_{i=1}^n$ $x_i \in \mathbb{R}$ and $y_i \in \{-1, +1\}$. Furthermore, assume that these samples are i.i.d. and for every i it holds that (x_i, y_i) is drawn according to the joint probability density function displayed in figure 1. You want to train a classifier for estimating the label y_{new} of a new point based on the feature x_{new}.

(j) [ ] Which of the following classifier should you use for this task? Explain your answer.

1. Logistic Regression 2. Linear Discriminant Analysis (LDA) 3. Gaussian Bayes Classifier

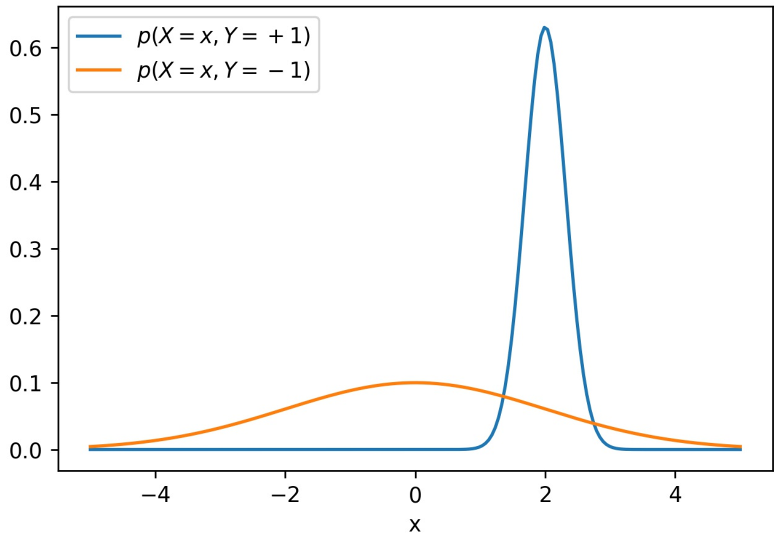

Figure 1: Joint probability density function of X and Y.

Exercise 2: Gaussian-Mixture Bayes Classifier

A Gaussian-Mixture Bayes Classifier is a generative model which explicitly models $P(Y = y)$ and $p_{X|Y}(x|y)$. The prior $P(Y = y)$ is modeled with

$$P(Y = y; p) = \text{Categorical}(y; p)$$

and the likelihood $p_{X|Y}(x|y)$ is modeled as a Gaussian Mixture with

$$p_{X|Y}(x|y; k, w, \mu, \Sigma) = \sum_{j=1}^{k} w_j^{(y)} \mathcal{N}(x; \mu_j^{(y)}, \sigma^2{}_j^{(y)}) \, .$$

p		Parameter of categorical distribution.
k		Number of clusters per class.
$w_j^{(y)}$	$\forall j \in \{1, \ldots, k\}, \forall y \in \{-1, +1\}$	$P($" New point lies in cluster j of class y"$\|Y = y)$
$\mu_j^{(y)}$	$\forall j \in \{1, \ldots, k\}, \forall y \in \{-1, +1\}$	Center of the cluster j of class y.
$\sigma^2{}_j^{(y)}$	$\forall j \in \{1, \ldots, k\}, \forall y \in \{-1, +1\}$	variance of the cluster j of class y.

For classification it computes

$$\begin{aligned} \arg\max_y P(Y = y|X = x) &= \arg\max_y P(Y = y|X = x) \cdot p_X(x) \\ &= \arg\max_y p_{X|Y}(x|y) \cdot P(Y = y). \end{aligned}$$

Suppose you got a data set $\{(x_i, y_i)\}_{i=1}^{10'000}$ $x_i \in \mathbb{R}$ and $y_i \in \{-1, +1\}$. Furthermore, you assume that the samples are drawn i.i.d.. In figure 2 the histogram of your data set is displayed. Your job is to train a classifier for estimating the label y_{new} of a new point based on the feature x_{new}.

In this situation ...

(a) [✓] ... one gets better results with a Gaussian-Mixture Bayes Classifier than with a Gaussian Bayes Classifier. ☐ True ☐ False

(b) [✓] ... one gets better results with a Gaussian Naive Bayes Classifier than with a Gaussian Bayes Classifier. ☐ True ☐ False

Now you decide to use a Gaussian-Mixture Bayes Classifier.

(c) [✓] How do you choose the value of k in this situation?

1. 1 2. 2 3. 3 4. 4 5. 5

(d) [✓] Suppose one chooses $k = 10$. Then the classification performance will decrease strongly. ☐ True ☐ False

(e) [✓] Suppose one chooses k higher than the number of samples. Then the classification performance will decrease strongly. ☐ True ☐ False

(f) Let p_{+1} be the parameter that models $P(Y = +1)$. Derive how to train p_{+1} with MLE?

(g) You decide to train the parameters $w_j^{(y)}$, $\mu_j^{(y)}$ and $\sigma^2{}_j^{(y)}$ with MLE too. Derive the resulting optimization problem. You don't have to solve it.

(h) Now you decide to train the parameters $w_j^{(y)}$, $\mu_j^{(y)}$ and $\Sigma_j^{(y)}$ with the Hard EM-algorithm. Explicitly derive the E-Step.

(i) What problem do you expect if one uses the Hard EM-algorithm in this situation? Will the Soft EM-algorithm help?

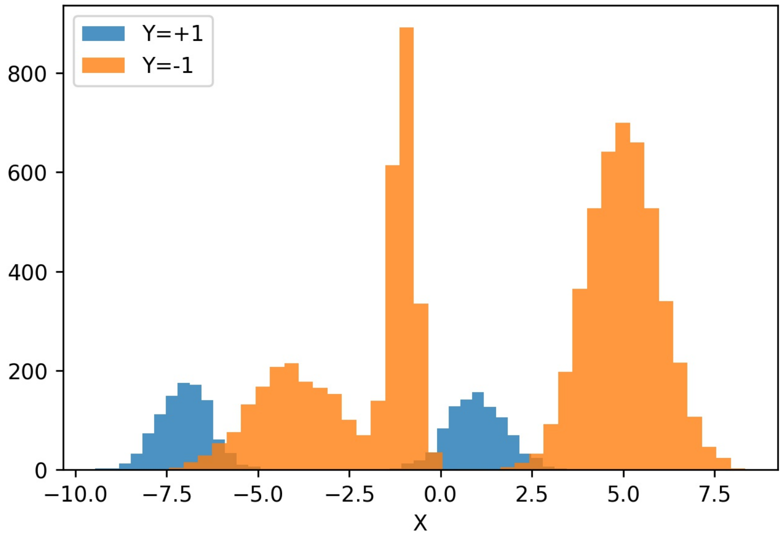

Figure 2: Histogram of the X-value of $10'000$ samples splitted according to their label.

Exercise 3: EM for mixture of distributions

We have a generative model for an integer random variable x over the values $1, 2, 3$. The generative model uses two distributions:

$$p_1(x) = \begin{cases} \alpha & if\ x = 1 \\ 1 - \alpha & if\ x = 2 \\ 0 & if\ x = 3 \end{cases} \tag{1}$$

$$p_2(x) = \begin{cases} 0 & if\ x = 1 \\ 1 - \beta & if\ x = 2 \\ \beta & if\ x = 3 \end{cases} \tag{2}$$

The overall generative models reads then $p(x) = \gamma p_1(x) + (1 - \gamma) p_2(x)$. The number of observations are respectively: $k_1, k_2, k_3 = (30, 20, 60)$. The EM is initialized with $\alpha_0, \beta_0, \gamma_0 = (\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$.

(a) Write the joint distribution over observed and latent variables governed by the parameters $\theta = (\alpha, \beta, \gamma)$ for a single sample x, z.

(b) E step. Evaluate the responsibilities using the current parameter values.

(c) M step. Re-estimate the parameters using the current responsibilities.

(d) Using given numbers of observations, calculate E and M steps until convergence.

Exercise 4: Generative adversarial networks

You train a generative adversarial network (GAN) with neural network discriminator D and neural network generator G. Let $\mathbf{z} \sim N(0, I)$, where I is the nxn identity matrix, represent the random Gaussian input for G. The objective during training is given by

$$\min_G \max_D \mathbb{E}_{\mathbf{x} \sim p_{data}}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z} \sim N(0,I)}[\log(1 - D(G(\mathbf{z})))], \tag{3}$$

where p_{data} is the data generating distribution. Answer the following questions:

(a) [ ] If D and G both have enough capacity, i.e., if they can model arbitrary functions, the optimal G will be such that:

1. $G(\mathbf{z}) \sim N(0, I)$ 2. $G(\mathbf{z}) \sim p_{data}$ 3. $G(\mathbf{z}) \sim p_{data} * N(0, I)$

where $*$ is the convolution symbol.

(b) [ ] The objective above can be interpreted as a two-player game between G and D.

1. True 2. False

(c) Suppose that the probability of a training sample $\mathbf{x}$ is $p_{data}(\mathbf{x}) = \frac{1}{100}$ and the probability of $\mathbf{x}$ under G is $p_G(\mathbf{x}) = \frac{1}{50}$. Suppose that the discriminator D is the globally optimal discriminator for G with the above loss. What is the probability of D classifying $\mathbf{x}$ as being from the generator?

Prove that, in general, with $p_{data}(\mathbf{x}) = a$ if $p_G(\mathbf{x}) = b$, this probability is equal to $\frac{b}{b+a}$.