Oleksandr Tarasov
(Matrikelnumer: 12310556)

Assignment 1

submitted to
Deep Learning [708219,708220,INP31875UF]

Supervisors
Özdenizci Ozan

Graz, October 2023

1 Maximum Likelihood Estimation

Consider a classification problem with two classes C_0 and C_1. For each class C_k, the samples come from a d-dimensional Gaussian distribution with mean vector μ_k and a covariance matrix $\Sigma_k = \sigma_k^2 I_d$, where I_d is the $d \times d$ identity matrix and $\sigma_k \in R^+$. probability of data point vector x conditioned on class k equals:

$$p(x|C_k) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_k|}} \exp\left(-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\right)$$

Hint: $|\Sigma_k|$ is the determinant of $\Sigma_k = \sigma_k^2 I_d$, and equals σ_k^{2d}.

Your training set consists of samples $X = \langle x^{(1)}, \ldots, x^{(n)} \rangle$, where the data points $x^(m) \in R^d$ i.i.d. You have the corresponding binary targets $t = \langle t^{(1)}, \ldots, t^{(n)} \rangle$ with $t^{(m)} \in \{0,1\}$, which indicates the class of the input sample (i.e., $t^{(m)} = 1$ indicates class C_1). You will fit a parameterized model for the data-generating distribution:

$$p(X,t|\theta) = p(t|\theta) \times p(X|t,\theta)$$

Your model includes a prior probability for the occurrence of each class, where class C_0 occurs with probability $P(C_0) = p_0$ and class C_1 occurs with probability $P(C_1) = 1 - p_0$. The parameters of your model are $\theta = \langle p_0, \mu_0, \mu_1, \sigma_0, \sigma_0 \rangle$.

1.1 a

Write the likelihood $p(x^{(m)}, t^{(m)}|\theta)$ of a single example $x^{(m)}, t^{(tm)}$. Accordingly, write the likelihood $p(X,t|\theta)$ of the whole training set X,t and then use this to derive the log-likelihood of the training set.

The probability of a data point conditioned on class:

$$\begin{aligned}
p(x^m|C_{t^m}) &= \mathcal{N}(x^m; \mu_{t^m}, \Sigma_{t^m}) = \\
&= \frac{1}{\sqrt{(2\pi)^d |\Sigma_{t^m}|}} \exp\left(-\frac{1}{2}(x-\mu_{t^m})^T \Sigma_{t^m}^{-1}(x-\mu_{t^m})\right) = \\
&= \frac{1}{\sqrt{(2\pi)^d |\sigma_{t^m}^{2d} I|}} \exp\left(-\frac{1}{2}(x-\mu_{t^m})^T (\sigma_{t^m}^2 I)^{-1}(x-\mu_{t^m})\right)
\end{aligned}$$

Considering the fact that we have class labels of the data and $t \in \{0,1\}$:

$$
\begin{aligned}
p(x^m, t^m = 0|\theta) &= p(t^m = 0|\theta)p(x^m|t^m = 0, \theta) \\
&= p_0 \times \mathscr{N}(x^m; \mu_0, \Sigma_0) \\
p(x^m, t^m = 1|\theta) &= p(t^m = 1|\theta)p(x^m|t^m = 1, \theta) \\
&= (1 - p_0) \times \mathscr{N}(x^m; \mu_1, \Sigma_1) \\
p(x^m, t^m|\theta) &= [p_0 \times \mathscr{N}(x^m; \mu_0, \Sigma_0)]^{(1-t^m)} [(1 - p_0) \times \mathscr{N}(x^m; \mu_1, \Sigma_1)]^{t^m}
\end{aligned}
$$

The likelihood of a single point (x^m, t^m):

$$\mathscr{L}(\theta) = p(x^m, t^m|\theta)$$

The likelihood of the whole dataset (X, t):

$$\mathscr{L}(\theta) = \prod_{m=1}^{N} p(x^m, t^m|\theta)$$

The negative log-likelihood(NLL):

$$
\begin{aligned}
NLL(\theta) &= -\sum_{m=1}^{N} \log p(x^m, t^m|\theta) = \\
&= -\sum_{m=1}^{N} [(1 - t^m)(\log p_0 + \log \mathscr{N}(x^m; \mu_0, \Sigma_0) + t^m(\log(1 - p_0) + \log \mathscr{N}(x^m; \mu_1, \Sigma_1)))]
\end{aligned}
$$

1.2 b

Derive the maximum likelihood estimate of μ1 for this model.

$$
\begin{aligned}
\hat{\theta}_{MLE} &= \underset{\theta}{argmin}\, NLL(\theta) \\
\frac{\partial NLL(\theta)}{\partial \mu_1} &= 0 \\
\frac{\partial \left[-\sum_{m=1}^{N} t^m \log \mathscr{N}(x^m; \mu_1, \Sigma_1)\right]}{\partial \mu_1} &= 0
\end{aligned}
$$

Let N_1 bet the number of data points from the first class($t^m = 1$):

$$\frac{\partial \left[-\sum_{m=1}^{N} t^m \log \mathscr{N}(x^m; \mu_1, \Sigma_1)\right]}{\partial \mu_1} = \frac{\partial \left[-\sum_{m=1}^{N_1} \log \mathscr{N}(x^m; \mu_1, \Sigma_1)\right]}{\partial \mu_1} = 0$$

Let‘s consider the derivative of the $\log \mathcal{N}(x^m; \mu_1, \Sigma_1)$ expression:

$$
\begin{aligned}
\frac{\partial \log \mathcal{N}(x^m; \mu_1, \Sigma_1)}{\partial \mu_1} &= \frac{\partial \left[-\frac{1}{2}\log((2\pi)^d|\Sigma_1|) - \frac{1}{2}(x^m-\mu_1)^T\Sigma_1^{-1}(x^m-\mu_1)\right]}{\partial \mu_1} \\
&= \frac{\partial \left[-\frac{1}{2}(x^m-\mu_1)^T\Sigma_1^{-1}(x^m-\mu_1)\right]}{\partial \mu_1} \qquad \left(z_m = (x^m-\mu_1), \frac{\partial z^m}{\partial \mu_1} = -I\right) \\
&= \frac{\partial}{\partial_m}\left(z_m^T\Sigma_1^{-1}z_m\right)\frac{\partial z^m}{\partial \mu_1} \\
&= -2\Sigma_1^{-1}z_m \\
&= -2\Sigma_1^{-1}(x^m-\mu_1)
\end{aligned}
$$

Then the whole expression:

$$
\begin{aligned}
\frac{\partial \left[-\sum_{m=1}^{N_1}\log \mathcal{N}(x^m; \mu_1, \Sigma_1)\right]}{\partial \mu_1} &= 0 \\
-\sum_{m=1}^{N_1}\left[-2\Sigma_1^{-1}(x^m-\mu_1)\right] &= 0 \\
2\Sigma_1^{-1}\sum_{m=1}^{N_1}x^m - 2N_1\Sigma_1^{-1}\mu_1 &= 0 \\
\mu_1 &= \frac{\sum_{m=1}^{N_1}x^m}{N_1}
\end{aligned}
$$

So the MLE estimation for the μ_1 is the mean of all data points that are related to C_1 class.

1.3 c

Derive the maximum-likelihood estimate of p_0 for this model.

$$
\begin{aligned}
\frac{\partial NLL(\theta)}{\partial p_0} &= 0 \\
\frac{\partial \left[-\sum_{m=1}^{N}(1-t^m)\log p_0 + t^m \log(1-p_0)\right]}{\partial p_0} &= 0 \\
-\sum_{m=1}^{N}\left[\frac{1-t^m}{p_0} - \frac{t^m}{1-p_0}\right] &= 0 \\
-\sum_{m=1}^{N}\left[1-p_0-t^m+p_0t^m-p_0t^m\right] &= 0 \\
\sum_{m=1}^{N} t^m + Np_0 - N = 0 \\
p_0 = 1 - \frac{\sum_{m=1}^{N} t^m}{N}
\end{aligned}
$$

1.4 d

Let's say we are interested in classifying samples by minimizing expected loss, where the loss matrix L will be expressed as:

$$
L = \begin{bmatrix} 0 & 20 \\ 1 & 0 \end{bmatrix}
$$

Firstly, using Bayes' rule, express $p(C_0|x)$ and $p(C_1|x)$ in terms of p_0. Then use these to derive an expression for the loss, for each possible classification outcome (i.e., correct C_0, correct C_1, false C_0, false C_1).

Let‘s first consider the Bayes‘ rule:

$$
\begin{aligned}
p(C_0|x) &= \frac{p(x|C_0)p(C_0)}{P(x)} \\
&= \frac{p(x|C_0)p(C_0)}{p(x|C_0)p(C_0) + p(x|C_1)p(C_1)} \\
&= \frac{p_0\mathcal{N}(x,\mu_0,\Sigma_0)}{p_0\mathcal{N}(x,\mu_0,\Sigma_0) + (1-p_0)\mathcal{N}(x,\mu_1,\Sigma_1)} \\
p(C_1|x) &= \frac{(1-p_0)\mathcal{N}(x,\mu_1,\Sigma_1)}{p_0\mathcal{N}(x,\mu_0,\Sigma_0) + (1-p_0)\mathcal{N}(x,\mu_1,\Sigma_1)}
\end{aligned}
$$

Let‘s define losses:

$$
\begin{aligned}
Loss(correct\ C_0) &= L_{1,1}P(C_0|x) = 0 \times P(C_0|x) \\
Loss(correct\ C_1) &= L_{2,2}P(C_1|x) = 0 \times P(C_1|x) \\
Loss(false\ C_0) &= L_{2,1}P(C_0|x) \\
&= 1 \times \frac{p_0\mathcal{N}(x,\mu_0,\Sigma_0)}{p_0\mathcal{N}(x,\mu_0,\Sigma_0) + (1-p_0)\mathcal{N}(x,\mu_1,\Sigma_1)} \\
Loss(false\ C_1) &= L_{1,2}P(C_1|x) \\
&= 20 \times \frac{(1-p_0)\mathcal{N}(x,\mu_1,\Sigma_1)}{p_0\mathcal{N}(x,\mu_0,\Sigma_0) + (1-p_0)\mathcal{N}(x,\mu_1,\Sigma_1)}
\end{aligned}
$$