Elements of Machine Learning
Exercise Sheet 2
Winter Term 2023/2024

William LaCroix - wila00001@stud.uni-saarland.de - 7038732
Philipp Hawlitschek - phha00002@stud.unisaarland.de - 7043167

Problem 2 T, 12 points. Linear Discriminate Analysis

1. [2pts]

In the derivation LDA we assume that input feature $x \in \mathbb{R}$ is continuous. Here, we consider $x \in \mathbb{N}_0$ to be a natural number and x follows a Poisson distribution. So, $p_k(x)$ is given by

$$p_k\left(x\right) = Pr\left(Y = k|X = x\right) = \frac{\pi_k f_k\left(x\right)}{\sum_{\iota=1}^{K} \pi_\iota f_\iota\left(x\right)}, \quad where \quad f_k\left(x\right) = \frac{\lambda_k^x e^{-\lambda_k}}{x!}.$$

Per class k, there is separate distribution parameter $\lambda_k > 0$. Derive the decision boundary for two classes k and l with respect to x.

To find the decision boundary wrt. x, we must set the probabilities for the two classes to be equal and solve for x where $f_k\left(x\right) = \frac{\lambda_k^x e^{-\lambda_k}}{x!}$. This equation represents the line at which the choice between classes k and l for x is 50:50:

$$p_k\left(x\right) = p_l\left(x\right)$$
$$Pr\left(Y = k|X = x\right) = Pr\left(Y = l|X = x\right)$$
$$\frac{\pi_k f_k\left(x\right)}{\sum_{\iota=1}^{K} \pi_\iota f_\iota\left(x\right)} = \frac{\pi_l f_l\left(x\right)}{\sum_{\iota=1}^{K} \pi_\iota f_\iota\left(x\right)}$$
$$\pi_k f_k\left(x\right) = \pi_l f_l\left(x\right)$$
$$\pi_k \frac{\lambda_k^x e^{-\lambda_k}}{x!} = \pi_l \frac{\lambda_l^x e^{-\lambda_l}}{x!}$$
$$log\left(\pi_k \frac{\lambda_k^x e^{-\lambda_k}}{x!}\right) = log\left(\pi_l \frac{\lambda_l^x e^{-\lambda_l}}{x!}\right)$$
$$log\left(\pi_k\right) + \log\left(\lambda_k^x\right) + log\left(e^{-\lambda_k}\right) - log\left(x!\right) = log\left(\pi_l\right) + log\left(\lambda_l^x\right) + log\left(e^{-\lambda_l}\right) - log\left(x!\right)$$
$$log\left(\pi_k\right) + x\log\left(\lambda_k\right) + log\left(e^{-\lambda_k}\right) = log\left(\pi_l\right) + xlog\left(\lambda_l\right) + log\left(e^{-\lambda_l}\right)$$
$$xlog\left(\lambda_k\right) - xlog\left(\lambda_l\right) = log\left(\pi_l\right) - log\left(\pi_k\right) - log\left(e^{-\lambda_k}\right) + log\left(e^{-\lambda_l}\right)$$
$$x\left(log\left(\lambda_k\right) - log\left(\lambda_l\right)\right) = log\left(\pi_l\right) - log\left(\pi_k\right) - log\left(e^{-\lambda_k}\right) + log\left(e^{-\lambda_l}\right)$$
$$xlog\left(\frac{\lambda_k}{\lambda_l}\right) = log\left(\frac{\pi_l e^{-\lambda_l}}{\pi_k e^{-\lambda_k}}\right)$$
$$xlog\left(\frac{\lambda_k}{\lambda_l}\right) = log\left(\frac{\pi_l e^{\lambda_k}}{\pi_k e^{\lambda_l}}\right)$$
$$x = \frac{log\left(\frac{\pi_l e^{\lambda_k}}{\pi_k e^{\lambda_l}}\right)}{log\left(\frac{\lambda_k}{\lambda_l}\right)}$$

2. [4pts]

In the lecture we dealt with finitely many classes $k \in [K]$. In the following assume there are infinitely many classes $k \in \mathbb{N}_0$ (here classes start 0) and k is Poisson-distributed $\pi(k) = \frac{\lambda^k e^{-\lambda}}{k!}$. Moreover, assume $x \in \mathbb{R}$ and $f_k(x) \sim \mathcal{N}(\mu_k, \sigma^2)$. Derive the decision boundary for two classes k and l with respect to x. Hint: Perform similar steps as in (1.), except that $\pi(k)$ follows a Poisson distribution .

Let's perform the same trick, this time with $\pi(k) = \frac{\lambda^k e^{-\lambda}}{k!}$, $\quad f_k(x) \sim \mathcal{N}(\mu_k, \sigma^2) \approx \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu_k)^2}{2\sigma^2}}$, solve for x

$$
\begin{aligned}
p_k(x) &= p_l(x) \\
Pr(Y = k|X = x) &= Pr(Y = l|X = x) \\
\frac{\pi_k f_k(x)}{\sum_{\iota=1}^{\infty} \pi_\iota f_\iota(x)} &= \frac{\pi_l f_l(x)}{\sum_{\iota=1}^{\infty} \pi_\iota f_\iota(x)} \\
\pi_k f_k(x) &= \pi_l f_l(x) \\
\frac{\lambda^k e^{-\lambda}}{k!} \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu_k)^2}{2\sigma^2}} &= \frac{\lambda^l e^{-\lambda}}{l!} \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu_l)^2}{2\sigma^2}} \\
\frac{\lambda^k}{k!} \cdot e^{-\frac{(x-\mu_k)^2}{2\sigma^2}} &= \frac{\lambda^l}{l!} \cdot e^{-\frac{(x-\mu_l)^2}{2\sigma^2}} \\
log\left(\frac{\lambda^k}{k!} \cdot e^{-\frac{(x-\mu_k)^2}{2\sigma^2}}\right) &= log\left(\frac{\lambda^l}{l!} \cdot e^{-\frac{(x-\mu_l)^2}{2\sigma^2}}\right) \\
log\left(\lambda^k\right) - log(k!) + log\left(e^{-\frac{(x-\mu_k)^2}{2\sigma^2}}\right) &= log\left(\lambda^l\right) - log(l!) + log\left(e^{-\frac{(x-\mu_l)^2}{2\sigma^2}}\right) \\
\frac{(x-\mu_l)^2}{2\sigma^2} - \frac{(x-\mu_k)^2}{2\sigma^2} &= log\left(\lambda^l\right) - log\left(\lambda^k\right) - log(l!) + log(k!) \\
\frac{(x-\mu_l)^2 - (x-\mu_k)^2}{2\sigma^2} &= log\left(\frac{\lambda^l k!}{\lambda^k l!}\right) \\
(x-\mu_l)^2 - (x-\mu_k)^2 &= 2\sigma^2 log\left(\frac{\lambda^l k!}{\lambda^k l!}\right) \\
x^2 - 2x\mu_l + \mu_l^2 - x^2 + 2x\mu_k - \mu_k^2 &= 2\sigma^2 log\left(\frac{\lambda^l k!}{\lambda^k l!}\right) \\
x &= \frac{\sigma^2 log\left(\frac{\lambda^l k!}{\lambda^k l!}\right) - \mu_l^2 + \mu_k^2}{(\mu_k - \mu_l)}
\end{aligned}
$$

3. [3pts]

Next, we consider the multivariate setting $x = (x_1, ..., x_m)^T$ with K classes, i.e. $p(k) = \pi_k$. Furthermore, we assume that all x_i are mutually independent, that is $f(x) = \prod_{i=1}^{m} f_i(x_i)$, even when conditioned on a certain class. We assume that the variance is the same across classes, while the means differ per class. Thus, per class each feature is distributed as $f_{ki}(x_i) \sim \mathcal{N}(\mu_{ki}, \sigma_i^2)$ Derive the density $f_k(x)$. State (not derive) the discriminant δ_x.

Given $f_{ki}(x_i) \sim \mathcal{N}(\mu_{ki}, \sigma_i^2) \approx \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x_i-\mu_{ki})^2}{2\sigma^2}}$, we can begin rewriting:

$$f_k(x) = \prod_{i=1}^{m} f_{ki}(x_i)$$

$$f_k(x) = \prod_{i=1}^{m} \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x_i-\mu_{ki})^2}{2\sigma^2}}$$

$$f_k(x) = \frac{1}{\left(\sigma\sqrt{2\pi}\right)^m} \cdot \prod_{i=1}^{m} e^{-\frac{(x_i-\mu_{ki})^2}{2\sigma^2}}$$

$$f_k(x) = \frac{exp\left(\sum_{i=1}^{m} - \frac{(x_i-\mu_{ki})^2}{2\sigma^2}\right)}{\left(\sigma\sqrt{2\pi}\right)^m}$$

$$f_k(x) = \frac{exp\left(-\frac{m}{2\sigma^2} \cdot \sum_{i=1}^{m} (x_i - \mu_{ki})^2\right)}{\left(\sigma\sqrt{2\pi}\right)^m}$$

The discriminant $\delta_k(x)$ can be expressed as the log of the density function times the prior probability, $\delta_k(x) = log(p(k) \cdot f_k(x))$, thus:

$$\delta_k(x) = log\left(\frac{\pi_k}{\left(\sigma\sqrt{2\pi}\right)^m}\right) - \frac{m}{2\sigma^2} \cdot \sum_{i=1}^{m} (x_i - \mu_{ki})^2$$

4. [2pts]

What are the assumptions of LDA regarding the data distribution and decision boundary. How do they compare to the assumptions of the other models presented in the lecture (k-NN, Logistic regression, QDA)?

- LDA assumes features are normally distributed, with equal covariance and a linear decision boundary.
- k-NN makes no assumptions about the distribution and covariance of the features, or the shape of the decision boundary
- Logistic regression makes no asumptions about the distribution of the features, but that the log odds of the response variable is a linear combination of the feature variables, and the decision boundary is linear.
- QDA assumes normal distribution of features, like LDA, but allows separate covaraince matrices per feature, and the decision boundary can be quadratic, not just linear.

5. [1pts]

For the multivariate setting, give a heuristic depending on covariance matrix Σ, when it is better to use LDA or QDA. Why can it be preferable to use LDA instead of QDA in that case?

It is better to use LDA when the covariance matrix Σ is the same across classes, use QDA when the assumption of equal covariance across classes does not hold. It can be preferable to use LDA when a simpler model is called for, either to prevent overfitting data, or due to computational concerns.