Elements of Machine Learning, WS 2023/2024
Jilles Vreeken and Krikamol Muandet
Classwork Sheet #2: *Classification*

Problem 1 (C, For Tutorials 20.11 and 21.11). **Logistic regression.**

1. For which types of response variables is it better to use logistic or linear regression?
2. Assume you want to predict how many cars cross an intersection depending on date and and daytime. Is it better to use logistic or linear regression? What problems occur?
3. Read Chapter 4.6.1 & 4.6.2 in the ISLP book. Is the Poisson regression model better suited for the problem described above?
4. Show that for the sigmoid function $\sigma(x) = \frac{e^x}{1+e^x}$, the following two equations hold:
   (i) $\sigma(-x) = 1 - \sigma(x)$
   (ii) $\sigma'(x) = \sigma(x)(1 - \sigma(x))$.

Solution.

1. Linear regression is better for continous targets and logistic regression for a discrete target with finitely many instantiations, which can be seen as different categories.
2. Linear regression is better suited, since, in theory at least, the number of cars crossing an intersection can not be easily bouded. Moreover, how many cars cross an intersection is a quantitative quantitative and not a qualitive.
   Problems:
   - Linear regression will predict, if the gradient is not zero, values below zero, which does not make sense in this example.
   - The variance is not the same for every time of the day, instead it will vary depending on the daytime.
   - The target is discrete, i.e. assumes values in $\mathbb{N}_0$, hence predicting values not in $\mathbb{N}_0$ does not make sense.
3. Yes, the model is more suitable, since our target corresponds to a counting variable.
4. *Proof.*
   (i) $1 - \sigma(x) = \frac{1+e^x}{1+e^x} - \frac{e^x}{1+e^x} = \frac{1}{1+e^x} = \frac{e^{-x}}{1+e^{-x}} = \sigma(-x)$
   (ii) $\frac{d\sigma(x)}{dx} = \frac{e^x}{(1+e^x)^2} = \sigma(x)\left(\frac{1}{1+e^x}\right) = \sigma(x)\left(\frac{e^{-x}}{1+e^{-x}}\right) = \sigma(x)\sigma(-x) = \sigma(x)(1-\sigma(x))$

□

Problem 2 (C, For Tutorials 27.11 and 28.11). **Linear Discriminant Analysis**

1. Show that for a symmetric matrix $A \in \mathbb{R}^{n\times n}$ and two vectors $u, v \in \mathbb{R}$ it holds that $x^T Av = v^T Ax$.

2. Given two independent normally distributed random variables $x_1 \sim \mathcal{N}(\mu_1, 1)$ and $x_2 \sim \mathcal{N}(\mu_2, 1)$. Show that the joint density $p(x_1, x_2)$ is again a density of a normal distribution.

3. Given a set of n independent observations $\mathbf{x} = x_1, ..., x_n$ from a normal distribution with unknown mean μ and variance σ^2. Derive μ that maximizes the likelihood $\mathcal{L}(\mu, \mathbf{x})$:

$$\arg\max_{\mu} \mathcal{L}(\mu, \mathbf{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right)$$

4. Compute log odds for multivariate LDA in the multiclass setting and show that the log-odds are linear in x. The log-odds for a class k and l are defined as $\log\left(\frac{\Pr(Y=k|X=x)}{\Pr(Y=l|X=x)}\right)$.

Solution.

1. *Proof.* $x^T Av = (x^T Av)^T = (Av)^T x = v^T A^T x = v^T Ax$, the last equation follows from A being symmetric. □

2. Since x_1 and x_2 are independent the joint factorizes as:

$$\begin{aligned} p(x_1, x_2) = p(x_1)p(x_2) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_1 - \mu_1)^2\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_2 - \mu_2)^2\right) \\ &= \frac{1}{2\pi} \exp\left(-\frac{1}{2}\left[(x_1 - \mu_1)^2 + (x_2 - \mu_2)^2\right]\right) \\ &= \frac{1}{2\pi} \exp\left(-\frac{1}{2}\left[(x - \mu)^T (x - \mu)\right]\right) \end{aligned}$$

, for $\mu = (\mu_1, \mu_2)$, Hence, $p(x_1, x_2) \sim \mathcal{N}(\mu, \mathbb{1}_2)$ is a density function of a normal distribution.

3. First, derive the log-likelihood:

$$\log \mathcal{L}(\mu, \mathbf{x}) = \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

Note, that log is monotonically increasing thus it does not change the arg max. Next, take the derivative w.r.t. μ and solve for its roots:

$$\frac{\partial}{\partial\mu} \log \mathcal{L}(\mu, \mathbf{x}) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} -2(x_i - \mu) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) \stackrel{!}{=} 0$$

Solving for μ yields:

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = 0 \Leftrightarrow \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \mu = 0 \Leftrightarrow \sum_{i=1}^{n} x_i - n\mu = 0 \Leftrightarrow \mu = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Moreover, $\frac{\partial^2}{\partial^2\mu} \mathcal{L}(\mu, \mathbf{x}) = -\frac{n}{\sigma^2} < 0$, thus $\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$ is a global maximum and therefore the solution to $\arg\max_{\mu} \mathcal{L}(\mu, \mathbf{x})$.

4. *Proof.*

$$\begin{aligned}
\log\left(\frac{\Pr(Y=k \mid X=x)}{\Pr(Y=l \mid X=x)}\right) =& \log\left(\frac{\pi_k f_k(x)}{\pi_l f_l(x)}\right) \\
=& \log\left(\frac{\pi_k \exp\left(-\frac{1}{2}\left(x-\mu_k\right)^T \boldsymbol{\Sigma}^{-1}\left(x-\mu_k\right)\right)}{\pi_l \exp\left(-\frac{1}{2}\left(x-\mu_l\right)^T \boldsymbol{\Sigma}^{-1}\left(x-\mu_l\right)\right)}\right) \\
=& \log\left(\frac{\pi_k}{\pi_l}\right) - \frac{1}{2}\left(x-\mu_k\right)^T \boldsymbol{\Sigma}^{-1}\left(x-\mu_k\right) \\
&+ \frac{1}{2}\left(x-\mu_l\right)^T \boldsymbol{\Sigma}^{-1}\left(x-\mu_l\right) \\
=& \log\left(\frac{\pi_k}{\pi_l}\right) - \frac{1}{2}\left(\mu_k+\mu_l\right)^T \boldsymbol{\Sigma}^{-1}\left(\mu_k-\mu_l\right) \\
&+ x^T \boldsymbol{\Sigma}^{-1}\left(\mu_k-\mu_l\right) \\
=& m^T x + b,
\end{aligned}$$

for $b := \log\left(\frac{\pi_k}{\pi_l}\right) - \frac{1}{2}\left(\mu_k+\mu_l\right)^T \boldsymbol{\Sigma}^{-1}\left(\mu_k-\mu_l\right)$ and $m := \left(\boldsymbol{\Sigma}^{-1}\left(\mu_k-\mu_l\right)\right)_j$, one can see that the log-odds are linear in x. □