Multinomial Logistic Regression. Softmax. Cross-Entropy

In case of two classes $y = \pm 1$, we introduced the following model:

$$p_w(x) = P(Y=1|\ X=x) = \sigma(w_0 + w_1 x_1 + \ldots + w_d x_d) = \frac{e^{w^T \tilde{x}}}{1 + e^{w^T \tilde{x}}}.$$

$$P(Y=-1|\ X=x) = 1 - p_w(x) = \frac{1}{1 + e^{w^T \tilde{x}}}.$$

The linear function can be obtained as

$$w^T \tilde{x} = \log \frac{P(Y=1|\ X=x)}{P(Y=-1|\ X=x)}.$$

It is called *logit* or *log-odds*. We could find the weights $w_0,\ w_1,\ \ldots,\ w_d$ maximizing probability

$$L\left(w; Y^{(1)} = y^{(1)}, \ldots, Y^{(N)} = y^{(N)} | X^{(1)} = x^{(1)}, \ldots, X^{(N)} = x^{(N)}\right) = \prod_{i=1}^{N} \left(p_w(x^{(i)})\right)^{[y^{(i)}=1]} \left(1 - p_w(x^{(i)})\right)^{[y^{(i)}=-1]}$$

that is equivalent to minimizing log-loss

$$Loss(w) = \sum_{i=1}^{N} \log\left(1 + e^{-y^{(i)} w^T \tilde{x}}\right) \to \min_{w}.$$

If we have a lot of data, we can use a different approach. We could use two linear models, one for each class/probability:

$$a_1(x) = w_{1,0} + w_{1,1} x_1 + \ldots + w_{1,d} x_d = w_1^T \tilde{x},$$
$$a_2(x) = w_{2,0} + w_{2,1} x_1 + \ldots + w_{2,d} x_d = w_2^T \tilde{x}.$$

We can convert these numbers into probabilities as follows:

$$P(Y=1|\ X=x) = p_{w_1}(x) = \frac{e^{w_1^T \tilde{x}}}{e^{w_1^T \tilde{x}} + e^{w_2^T \tilde{x}}}, \qquad P(Y=-1|\ X=x) = p_{w_2}(x) = \frac{e^{w_2^T \tilde{x}}}{e^{w_1^T \tilde{x}} + e^{w_2^T \tilde{x}}}.$$

Problem 1. For a given x consider the following random variables:

$Y^{(i)} \vert X^{(i)} = x^{(i)}$	-1	1
P	$p_{w_2}(x^{(i)})$	$p_{w_1}(x^{(i)})$

Write likelihood function $L(w_1, w_2)$ and log-loss used to find $w_{1,0}, w_{1,1}, \ldots, w_{1,d}, w_{2,0}, w_{2,1}, \ldots, w_{2,d}$.

Consider data of C classes, i.e., $y \in \{1, 2, \ldots, C\}$. Introduce C linear models:

$$\begin{array}{c} a_1(x) = w_{1,0} + w_{1,1}x_1 + w_{1,2}x_2 + \ldots + w_{1,d}x_d = w_1^T \tilde{x}, \\ a_2(x) = w_{2,0} + w_{2,1}x_1 + w_{2,2}x_2 + \ldots + w_{2,d}x_d = w_2^T \tilde{x}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_C(x) = w_{C,0} + w_{C,1}x_1 + w_{C,2}x_2 + \ldots + w_{C,d}x_d = w_C^T \tilde{x}. \end{array}$$

Problem 2. Check that the following numbers (called *softmax*) sum up to one:

$$p_{w_1}(x) = \frac{e^{a_1(x)}}{\sum\limits_{k=1}^{C} e^{a_k(x)}}, \quad \ldots, \quad p_{w_C}(x) = \frac{e^{a_C(x)}}{\sum\limits_{k=1}^{C} e^{a_k(x)}}.$$

Problem 3. For a given $x^{(i)}$ consider the following random variable:

$Y^{(i)} \mid X^{(i)} = x^{(i)}$	1	2	$\ldots$	C
P	$p_{w_1}(x^{(i)})$	$p_{w_2}(x^{(i)})$	$\ldots$	$p_{w_C}(x^{(i)})$

Write likelihood function $L(w_1, w_2, \ldots, w_C)$ and log-loss used to determine the parameters.

Problem 4. Show that the loss function from Problem 3 is a cross-entropy between random variable $Y^{(i)}|x^{(i)}$ and constant random variable

$Z^{(i)} \mid X^{(i)} = x^{(i)}$	1	2	$\ldots$	$y^{(i)}$	$\ldots$	C
P	0	0	$\ldots$	1	$\ldots$	0

In other words, $Z^{(i)}|X^{(i)} = x^{(i)}$ is one-hot encoder of the true target $y^{(i)}$ for the point $x^{(i)}$.

Hint: Cross-entropy

$$H(Z^{(i)}|x^{(i)}, Y^{(i)}|x^{(i)}) = \mathrm{P}(Z^{(i)} = 1|x^{(i)}) \log p_{w_1}(x^{(i)}) + \ldots + \mathrm{P}(Z^{(i)} = C|x^{(i)}) \log p_{w_C}(x^{(i)}) = \log p_{w_{y^{(i)}}}(x^{(i)}).$$