1 Logistic Regression

Logistic Regression is a problem known from machine learning and statistics. The task is to find a parameterized model $f_\theta(x)$ which assigns a class label to a given point, e.g., given the health record of a patient, assign whether he or she suffers from dementia. The model is chosen such that it minimizes the true risk, the probability of assigning the wrong class label (classifying healthy patients as having dementia or vice-versa) over the distribution of data points. Unfortunately, we can not compute the true risk, as it requires the knowledge over an infinite number of data points. In practice, we are given a finite set of data points $D = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(\ell)}\ell, y^{(\ell)})\}$, $x^{(i)} \in \mathbb{R}^d$, $y^{(i)} \in \{-1, 1\}$ where the true class labels are known. We assign the labels from the model by taking its sign: $\hat{y} = \operatorname{sign}(f_\theta(x))$. Now, we can compute an *empirical* estimate of the true risk

$$L^{\text{emp}}(\theta) = \frac{1}{\ell}\sum_{i=1}^{\ell}[\operatorname{sign}(f_\theta(x^{(i)})) \neq y^{(i)}] \ .$$

Here, $[\operatorname{sign}(f(y)) \neq y]$ is 1 if the sign of $f(x)$ does not agree with the label y, 0 otherwise.

The empirical risk is computed based on binary decisions: did the model assign the right class label or not? Therefore, the empircial risk is not continuous, not differentiable and can not be optimized easily. In fact, optimizing it is NP-hard under almost all circumstances. Instead we use a *surrogate* loss function from which we can hope that it will find a model which is close to the model optimizing the empirical risk.

We can transform the output of the model to a probability of the true class-label via the sigmoid function

$$\sigma(x) = \frac{1}{1 + \exp(-x)} \ .$$

This function is monotonous increasing in x, takes values in the range $[0, 1]$ and it holds $1 - \sigma(x) = \sigma(-x)$. With this, the probability of assigning the correct label to a given point x can be modeled as

$$P\left(Y = y \mid X = x\right) = \sigma(y \cdot f_\theta(x)) \ .$$

This leads us to our minimization problem: minimize the negative log-probability of the dataset

$$\begin{aligned} L(\theta) &= -\log \prod_{i=1}^{\ell} P\left(Y = y^{(i)} \mid X = x^{(i)}\right) \\ &= -\log \prod_{i=1}^{\ell} \sigma(y^{(i)} \cdot f_\theta(x^{(i)})) \\ &= \sum_{i=1}^{\ell} \log(1 + \exp(-y^{(i)} \cdot f_\theta(x^{(i)}))) \ . \end{aligned}$$

The logarithm does not change the optimum as it is a monotonous increasing function. However it makes all calculations and computations much easier.

So far, we have not discssed the choice of our model, $f_\theta(x)$. In theory, we are free to choose an arbitrary function here, but for logistic regression a linear model is used:

$$f_\theta(x) = \theta_0 + \sum_{j=1}^{d} \theta_j x_j.$$

2 Regularized Logistic Regression

If the number of data points ℓ is small, we can not hope to find the true parameter vector θ. Especially, if $\ell < d$ we will always find a setting of the parameters which fits the dataset optimally. However, this does not mean that the parameters will work well on future data points. This is especially a problem in gene-sequence analysis as gene-sequences are made up of millions of genes while typically only a few thousand sequences are available. A technique called regularization can help here. We augment the loss function by another term, which punishes weight vectors which lead to *complex* models. Two important choices are two-norm regularisation

$$L^2(\theta) = L(\theta) + \lambda \sum_{j=1}^{d} \theta_j^2 \ ,$$

and one-norm regularization

$$L^1(\theta) = L(\theta) + \lambda \sum_{i=j}^{d} |\theta_j|$$

One-norm regularization is especially important for gene-sequence analysis as it forces parameter values down to zero, while two-norm regularization only makes the norm of the parameter vector small. However, optimizing the one-norm directly is difficult as its derivative is not continuous: it is 1 for $\theta_j > 0$ and -1 if $\theta_j < 0$, leading to discontinuity at $\theta_j = 0$. A solution we will pursue in this course is reformulating the optimization problem so that we replace the discontinuity in the derivative by a pair of constraints. We can reformulate $\theta_j = w_j^+ - w_j^-$ with $w_j^+ \geq 0$ and $w_j^- \geq 0$ for $j = 1, \dots, d$. We obtain:

$$\begin{aligned}
&L(\theta) + \lambda \sum_{j=1}^{d} |\theta_j| \\
=&L((\theta_0, w_1^+ - w_1^-, \dots, w_d^+ - w_d^-)) + \lambda \sum_{j=1}^{d} |w_j^+ - w_j^-| \\
\leq&L((\theta_0, w_1^+ - w_1^-, \dots, w_d^+ - w_d^-)) + \lambda \sum_{j=1}^{d} w_j^+ + \lambda \sum_{i=1}^{d} w_j^-
\end{aligned}$$

In the last step we used the inequality $|a - b| \leq |a| + |b|$ and that all w are positive. The inequality is exact at the optimum: each pair w_j^+ and w_j^- will have at least one zero, provided we enforce the simple set of inequality constraints $w_j^+ \geq 0$ and $w_j^- \geq 0$ for $j = 1, \dots, d$. Taking the derivative of the expression above is easy and we have no longer a discontinuity.