ICTS 4156: Introduction to ML

Razvan Bunescu

Lecture notes, April 7, 2021

1 Notes on Quiz

1.1 Kernel notes

K_1 is a valid kernel. By definition, $\exists \phi_1$ such that $K_1(x,y) = \phi_1(x)^T \phi_1(y)$.

Let $K(x,y) = cK_1(x,y) = c\phi_1(x)^T\phi_1(y) = (\sqrt{c}\phi_1(x))^T(\sqrt{c}\phi_1(y)) = \phi(x)^T\phi(y)$, where $\phi(x) = \sqrt{c}\phi_1(x)$. Therefore, because $K(x,y) = \phi(x)^T\phi(y)$, this means K is a valid kernel.

2 Notes on Homework

2.1 Relative error reduction on Spam classification

Relative error reduction $\dfrac{a_2 - a_1}{1 - a_1} = \dfrac{0.981 - 0.978}{1 - 0.978} = \dfrac{0.003}{0.022} = 13.6\%$ relative error reduction.

3 Notes on Lecture Material

3.1 Binary Logistic Regression

$p(C_1|x) = \sigma(w^Tx) = \dfrac{1}{1 + e^{-w^Tx}}$

At test time, we predict positive class C_1 if and only if $p(C_1|x) > P(C_2|x)$. We know that $p(C_1|x) + P(C_2|x) = 1$, which means that $p(C_1|x) > P(C_2|x) => P(C_1|x) > 0.5$.

This means that $\dfrac{1}{1 + e^{-w^Tx}} > 0.5 = \dfrac{1}{2}$, which means that $e^{-w^Tx} < 1$, therefore $-w^Tx < 0$, which means $z = w^Tx > 0$.

Training Logistic Regression parameters w will be done by using the Maximum Likelihood Estimation principle, which says that we want to select the parameters that maximize the probability of the true labels:

$$\hat{w} = \arg\max_w P(t_1, t_2, ..., t_N|w) \tag{1}$$

The probability $P(t_n|w)$ is equal with:

- $P(t_n|w) = \sigma(w^T x_n) = h_n$, if $t_n = 1$.
- $P(t_n|w) = 1 - \sigma(w^T x_n) = 1 - h_n$, if $t_n = 0$.

Can we show that $P(t_n|w) = h_n^{t_n}(1 - h_n)^{(1-t_n)}$?

- If $t_n = 1$, we get $P(t_n|w) = h_n^{t_n}(1 - h_n)^{(1-t_n)} = h_n^1(1 - h_n)^0 = h_n$. Verified!
- If $t_n = 0$, we get $P(t_n|w) = h_n^{t_n}(1 - h_n)^{(1-t_n)} = h_n^0(1 - h_n)^1 = 1 - h_n$. Verified!

Assume the training examples are **independent identically distributed (i.i.d)**. The **likelihood** function:

$$
\begin{aligned}
P(t_1, t_2, ..., t_N|w) &= \prod_{n=1}^{N} && (2)\\
P(t_1, t_2, ..., t_N|w) &= \prod_{n=1}^{N} h_n^{t_n}(1 - h_n)^{(1-t_n)} && (3)\\
P(t_1, t_2, ..., t_N|w) &= \prod_{n=1}^{N} \sigma(w^T x_n)^{t_n}(1 - \sigma(w^T x_n))^{(1-t_n)} && (4)\\
& && (5)
\end{aligned}
$$

Mathematically, the weight vector w that maximizes the likelihood is going to be the same as the weight vector that maximizes the **log likelihood**. But this is equivalent with finding the weight vector w that minimizes the **negative log-likelihood**:

$$
\begin{aligned}
-\ln P(t_1, t_2, ..., t_N|w) &= -\sum_{n=1}^{N} t_n \ln h_n + (1 - t_n) \ln(1 - h_n) && (6)\\
-\ln P(t_1, t_2, ..., t_N|w) &= -\sum_{n=1}^{N} t_n \ln \sigma(w^T x_n) + (1 - t_n) \ln(1 - \sigma(w^T x_n)) && (7)
\end{aligned}
$$