HW Assignment 3 (Due by 10:30am on Oct 12)

1 Theory (100 points)

1. [**Maximum Likelihood**, 20 points]
   The Poisson distribution specifies the probability of observing k events in an interval, as follows:
   $$P(k \text{ events in interval}) = e^{-\lambda}\frac{\lambda^k}{k!} \tag{1}$$
   For example, k can be the number of meteors greater than 1 meter diameter that strike Earth in a year, or the number of patients arriving in an emergency room between 10 and 11 pm[1].

   Suppose we observe N samples k_1, k_2, ..., k_N from this distribution (i.e. numbers of meteors that strike Earth over a period of N years). Derive the maximum likelihood estimate of the event rate λ.

2. [**Logistic Regression**, 20 points]
   Consider a dataset that contains the 4 examples below i.e., the truth table of the logical XOR function. Prove that no logistic regression model can perfectly classify this dataset. Do not forget the bias feature $x_0 = 1$.

x_1	x_2	t
0	0	0
0	1	1
1	0	1
1	1	0

   Hint: Prove that there cannot be a vector of parameters $\mathbf{w}$ such that $P(t = 1|\mathbf{x}, \mathbf{w}) \geq 0.5$ for all examples $\mathbf{x}$ that are positive, and $P(t = 1|\mathbf{x}, \mathbf{w}) < 0.5$ for all examples $\mathbf{x}$ that are negative.

3. [**Logistic Regression**, 20 points]
   Prove that the gradient (with respect to $\mathbf{w}$) of the negative log-likelihood error function for logistic regression corresponds to the formula shown in lecture 4:
   $$\nabla_{\mathbf{w}} E(\mathbf{w}) = \sum_{n=1}^{N} (h_n - t_n)\mathbf{x}_n \tag{2}$$

4. [**Logistic Regression**, 20 points]
   In `scikit`, the objective function for logistic regression expresses the trade-off between training error and model complexity through a parameter C that is multiplied with the error term, as shown below. See the `scikit` documentation at `http://scikit-learn.org/stable/modules/linear_model.html#logistic-regression`.
   $$E(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C * \sum_{n=1}^{N} \ln(e^{-t_n(\mathbf{w}^T\mathbf{x}_n)} + 1) \tag{3}$$

[1] https://en.wikipedia.org/wiki/Poisson_distribution

- Show that the sum in the second term is equal with the negative log-likelihood, where $t_n = +1$ stands for positive labels and $t_n = -1$ stands for negative labels.
- Compute the C parameter such that the objective is equivalent with the standard formulation shown on the slides in which the regularization parameter λ is multiplied with the L2 norm term.

5. [**Softmax Regression**, 20 points]
Show that Logistic Regression is a special case of Softmax Regression. That is to say, if $\mathbf{w_1}$ and $\mathbf{w_2}$ are the parameter vectors of a Softmax Regression model for the case of two classes, then there exists a parameter vector $\mathbf{w}$ for Logistic Regression that results in the same classification as the Softmax Regression model.

6. [**Softmax Regression (*)**, 20 points]
Prove that the gradient (with respect to $\mathbf{w}_k$) of the negative log-likelihood error function for regularized softmax regression corresponds to the formula shown in lecture 4, for any class $k \in [1..K]$:

$$\nabla_{\mathbf{w}_k} E(\mathbf{w}) = -\frac{1}{N}\sum_{n=1}^{N}(\delta_k(t_n) - p(C_k|\mathbf{x}_n))\mathbf{x}_n + \alpha \mathbf{w}_k \tag{4}$$

2 Submission

Turn in a hard copy of your homework report at the beginning of class on the due date. On this theory assignment, **clear and complete explanations and proofs of your results are as important as getting the right answer**.