HW Assignment 9 (Due by 5:30 pm on May 5)

1 Theory (250 points)

1. [**Maximum Likelihood**, 30 points]
The Poisson distribution specifies the probability of observing k events in an interval, as follows:

$$P(k \text{ events in interval}) = e^{-\lambda}\frac{\lambda^k}{k!} \quad (1)$$

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 independent 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, i.e. the value that maximizes the likelihood (the probability of seeing these samples). This is similar to what we did in class for finding the parameters of logistic regression, e.g. formulate the likelihood, take the log, and find the parameters that maximize the log-likelihood. Another example is provided in Section 2.1 here: `http://www.cs.cmu.edu/~tom/mlbook/Joint_MLE_MAP.pdf`.

2. [**Logistic Regression**, 40 points]
Consider a training set that contains the following 8 examples: Prove that no binary

$\mathbf{x}$	x_1	x_2	x_3	$t(x)$
$\mathbf{x}^{(1)}$	0	0	0	+1
$\mathbf{x}^{(2)}$	0	1	0	+1
$\mathbf{x}^{(3)}$	1.5	0	-1.5	+1
$\mathbf{x}^{(4)}$	1.5	1	-1.5	+1
$\mathbf{x}^{(5)}$	1.5	0	0	–1
$\mathbf{x}^{(6)}$	1.5	1	0	–1
$\mathbf{x}^{(7)}$	0	0	-1.5	–1
$\mathbf{x}^{(8)}$	0	1	-1.5	–1

logistic regression model can perfectly classify this dataset. Do not forget to include the bias feature $x_0 = 1$.

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. [**Perceptrons**, 20 points]
A kernel perceptron for binary classification is run for a number of epochs E on a training dataset containing N examples, resulting in the dual parameters α_1, α_2, ..., α_N. What is the total number of mistakes that are made during training?

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

4. [**Logistic Regression**, 30 points]
In `scikit`, the objective function for binary 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`.

$$E(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C * \sum_{n=1}^{N} \ln(e^{-t_n(\mathbf{w}^T\mathbf{x}_n)} + 1) \tag{2}$$

- Show that the sum in the second term is equal with the negative log-likelihood.
- 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 binary 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 binary Logistic Regression that results in the same classification as the Softmax Regression model.

6. [**Kernel Techniques**, 30 points]
(a) Show that if $k_1(\mathbf{x}, \mathbf{y})$ and $k_2(\mathbf{x}, \mathbf{y})$ are valid kernel functions, then $k(\mathbf{x}, \mathbf{y}) = k_1(\mathbf{x}, \mathbf{y}) + k_2(\mathbf{x}, \mathbf{y})$ is also a valid kernel.
(b) Let $K : X \times X \to R$ be a kernel function defined over a sample space X. Prove that the function below (a *normalized kernel*) is a valid kernel.

$$\frac{K(x, y)}{\sqrt{K(x, x)K(y, y)}}$$

(c) Why it would not make sense to normalize a Gaussian kernel?

7. [**Naive Bayes**, 40 points]
The Naive Bayes algorithm for text categorization presented in class treats all sections of a document equally, ignoring the fact that words in the title are often more important than words in the text in determining the document category. Describe how you would modify the Naive Bayes algorithm for text categorization to reflect the constraint that words in the title are K times more important than the other words in the document for deciding the category, where K is an input parameter (include pseudocode).

8. [**Kernel Nearest Neighbor**, 20 points]
The nearest-neighbour classifier 1-NN assigns a new input vector $\mathbf{x}$ to the same class as that of the nearest input vector $\mathbf{x}_n$ from the training set, where in the simplest case, the distance is defined by the Euclidean metric $\|\mathbf{x} - \mathbf{x}_n\|^2$. By expressing this rule in terms of dot products and then making use of kernel substitution, formulate the nearest-neighbour classifier for a general nonlinear kernel.

9. [**Distance-Weighted Nearest Neighbor**, 20 points]
We have seen how to use kernels to formulate a distance-weighted nearest neighbor algorithm, when the labels are binary. Formulate a kernel-based, distance-weighted nearest neighbor that works for K classes, where $K \geq 2$.

2 Submission

Submit on Canvas a **PDF** of your homework report by the due date. On this theory assignment, **clear and complete explanations and proofs of your results are as important as getting the right answer**.