Advanced Techniques in Machine Learning (89-654-01)
Exam - Moed Aleph

Dr. Yossi Keshet
July 29, 2015

Instructions
- Total time: 3 hours.
- All written or printed material is allowed.
- You have to answer on all of the 3 questions below.
- Every item in every question weights 10 points.
- Explain your steps.
- Good luck!

1. Let \( S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \) be a training set of examples, where \( x_i \in \mathbb{R}^d \) is an input feature vector of length \( d \) and \( y_i \in \{1, \ldots, k\} \) is the output, where \( k \) is the number of classes. Define the logistic regression conditional probability \( \hat{P}(Y = y | X = x) \) for multiclass classification as
\[
\hat{P}(Y = y | X = x) = \frac{e^{w_y \cdot x}}{\sum_{r=1}^{k} e^{w_r \cdot x}},
\]
where \( \{w^1, \ldots, w^k\} \) are the parameters, such that \( w^r \) is the weight vector of class \( r \).

(a) Express the optimization problem that minimizes the regularized minus log likelihood over the training set.

(b) Assume that a cost function \( \ell(y, y') \) between two labels \( y \) and \( y' \) is given. It is a function that returns some positive number for every pair \( y \) and \( y' \) and returns 0 if \( y = y' \). For example, if \( y = 3 \) and \( y' = 2 \) the cost is not big, say 0.1, if \( y = 3 \) and \( y' = 8 \), this is unbearable and the cost is 1, and if \( y = 3 \) and \( y' = 3 \) then the cost is 0. Our goal is to propose an algorithm that minimizes the expected cost \( \mathbb{E}_{(x, y) \sim \rho}[\ell(y, \hat{y}(x))] \), where \( \hat{y} \) is the prediction of the algorithm. What can you say about the optimization problem in (a) regarding minimization of the cost function?

(c) Instead of maximizes the log likelihood, a new surrogate loss function is proposed to minimize the expected cost:
\[
\tilde{\ell}(\{w^1, \ldots, w^k\}, x, y) = \sum_{\tilde{y}=1}^{k} \hat{P}(Y = \tilde{y} | X = x) \ell(y, \tilde{y})
\]
Write the new optimization problem.
(d) Derive an algorithm using stochastic gradient descent method for the new optimization problem.

2. There are 20 different amino acids denoted by a single letter: A, C, D, E, F, ... The amino acids are the building blocks of proteins. Different proteins are made up of different combinations of amino acids. We would like to build a search engine for proteins, given a sub-sequence of amino acids. For example, the query for the search engine might be \{M,K,W,V,T\}, the search engine will result with all the proteins that are composed with the sub sequence \{M,K,W,V,T\}, for example Albumin, which is composed of \{M,K,W,V,T,F,L,L,L,L\}, but also of other proteins.

(a) Let \(a \in \Sigma\) be an amino acid, where \(\Sigma = \{A,C,D,E,F,\ldots\}\) is the set of all 20 amino acids. Let \(p \in \Sigma^*\) be a protein, defined as a sequence of amino acids. Design a function \(\phi(\bar{a}, p)\) that given a sub-sequence \(\bar{a} \in \Sigma^*\) of amino acids and a protein, will have high value if \(\bar{a}\) is in \(p\). Note that the order of the amino acids in \(\bar{a}\) or in \(p\) does not matter.

(b) We have a training set \(S\) of \(m\) example, \(S = \{(\bar{a}^i, (p^i_1, \ldots, p^i_N), y^i)\}_{i=1}^m\), where \(\bar{a}^i\) is the query to the search engine, \(p^i_j\) is a protein that is build of \(\bar{a}^i\), and its relevance is \(y^i\). \(y^i \in \{0,1,2,3,4,5\}\) and 0 means non-relevant. Propose an algorithm for ranking the proteins given a sequence of amino acids. Use the function \(\phi(\bar{a}, p)\).

(c) In January 2015 a new protein was found. We have the unique structure of the protein and we know its sequence of amino acids. How should we change \(\phi(\bar{a}, p)\) and how should we adapt the algorithm in order to include the new protein.

3. Structured prediction. We would like to predict an event in time - see figure 1. The input signal is divided into \(T\) frames (in the figure \(T=1000\)), and \(T\) is different for each and every example. At each frame we extract a feature vector. At the \(t\)-th frame the feature vector \(\vec{x}_t\) is composed of different measures of energy, the frequency of the sinusoid in the signal, and so on. We are given a training set of \(m\) examples. Each example is composed of a sequence of feature vector, \(\vec{x} = (x_1, \ldots, x_T)\), and a label \(\vec{y} = (t_b, t_e)\), where \(0 < t_b < t_e < T\). The input signal might have small noise of short time as depicted in the figure. Our goal is to design a learning algorithm that finds such events in time.

(a) How can you use a binary classifier to solve the problem? what might be the problem with such algorithm?

(b) We would like to have a structured prediction algorithm that take advantage of the special shape of the signal, its typical duration and so on. Design 3 feature function \(\phi(x, y)\) that are correlated with the typical signal.

(c) Give an example of a cost function between the two labels \(\ell(y, \hat{y})\) that ignore errors of \(t_b\) and \(t_e\) less than 10 msec, and propose an algorithm that predicts \(\hat{y}\) using the feature functions and aims at minimizing \(\ell(y, \hat{y})\) given a training set of examples.
Figure 1: Event in time \((t_b, t_e)\) and some noise.