UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
Department of Electrical and Computer Engineering

ECE 544NA Pattern Recognition
Fall 2014

EXAM 2 SOLUTIONS

Thursday, November 6, 2014

Problem 1 (34 points)

(a) Choose a direction, uniformly at random. Lean in that direction.

(b) If the ground slopes downhill, take a step.

(c) If the ground slopes uphill, take a step with probability $P(z_{k-1}, z_k)$ dependent on the current altitude z_{k-1} and proposed future altitude z_k, where

$$P(z_{k-1}, z_k) = e^{-\alpha(z_k - z_{k-1})/T_k} \tag{1}$$

The altitudes z_{k-1} and z_k are measured in meters above sea level; as you may know, altitudes in Urbana vary in the range $200 \leq z \leq 250$ meters (Amanda's house is at $z^* = 200$ meters). The variable T_k represents the tempo, in beats per minute, of the music on her mp3 player; since her battery is running out, this is a strictly non-increasing function of time, $T_{k-1} \geq T_k$. The constant $\alpha = 10$ bpm/meter.

(a) Given the information that you have, what are the smallest possible values of T_k that are guaranteed, with probability one, to eventually deliver Amanda to her house?

SOLUTION: $T_k = \frac{500}{\ln(k+1)}$

(b) Suppose that Amanda starts her journey in a bar at altitude $z_0 = 240m$. Is this information sufficient to change your answer to part (a)? Why or why not?

SOLUTION: No, because the bar might be clinging to the side of the hill in a valley 49.999 meters deep. Amanda's initial trajectory might take her to the bottom of that valley, and then, unless $T_k \geq \frac{500}{\ln(k+1)}$, she would not be guaranteed to recover.

(c) Suppose that now Amanda has developed the ability to teleport. Therefore her algorithm for getting home is now, for $k \in \{1, 2, \ldots\}$,

(i) Choose a target location somewhere in Urbana, uniformly at random.

(ii) If that location is downhill, teleport there.

(iii) If that location is uphill, teleport there with probability $P(z_{k-1}, z_k)$ given by Eq. 1.

Under this algorithm, what are the smallest possible values of T_k that guarantee, with probability one, that Amanda will eventually find her way home?

SOLUTION: All locations are now neighbors, therefore Amanda never has to travel uphill ever again, so $T_k = 0$ suffices.

(d) Suppose that we fix T_k at a constant slow tempo, $T_k = 2$bpm, for all time. Prove that, under this circumstance, Amanda is guaranteed to eventually reach home with probability one (although her algorithm will not guarantee that she stays there once she has arrived).

SOLUTION: Every location in Urbana can be reached, with nonzero probability, from any of its neighbors. There is therefore a path with nonzero probability from any starting location to any ending location. Since every location can be reached with nonzero probability from every other location, therefore with probability 1, every location will be reached eventually.

Problem 2 (33 points)

Consider an RBM with visible variables $v \in \{0,1\}^p$, and with hidden variables $h \in \Re^q$. Suppose that the joint probability $p(h,v)$ is given by

$$p(h,v) = \frac{1}{Z} e^{-h^T W v - \frac{1}{2} h^T R h} \tag{2}$$

where R is positive definite, and $W = [w_1, \ldots, w_p]$ for column vectors $w_k = [w_{1k}, \ldots, w_{qk}]^T$.

(a) Suppose that, for some odd reason, you have a training database in which both h_i and v_i are specified. Define

$$\mathcal{L} = \sum_{i=1}^{n} \ln p(h_i, v_i)$$

Compute $\partial\mathcal{L}/\partial w_{jk}$ as a function of $h_1, \ldots, h_n$, $v_1, \ldots, v_n$, W, and R. Your answer may include unevaluated explicit integrals or sums, but should not include any unevaluated expectations.

SOLUTION: $\frac{\partial\mathcal{L}}{\partial w_{jk}} = -\sum_{i=1}^{n} h_{ij} v_{ik} + n \frac{\sum_{v\in\{0,1\}^p} \int h_j v_k e^{-h^T W v} dh}{\sum_{v\in\{0,1\}^p} \int e^{-h^T W v} dh}$

(b) What is $E[h|v]$? Remember that $h \in \Re^q$. Specify your answer in terms of v, W, and R. Your answer should *not* include any unevaluated integrals, sums, or expectations. You may find it useful to know that every positive definite matrix, A, has a square root ($A = A^{1/2}A^{1/2}$), and an inverse ($AA^{-1} = I$), and that both $A^{1/2}$ and A^{-1} are also positive definite.

SOLUTION: $E[h|v] = -R^{-1}Wv$

(c) Now imagine the converse: you are given h. Remember that $v \in \{0,1\}^p$. Please compute the conditional probability that the j^{th} component of v is "turned on," that is, compute $\Pr\{v_j = 1|h\}$. Your answer should *not* include any unevaluated integrals, sums, or expectations.

SOLUTION: $\Pr\{v_j = 1|h\} = \frac{1}{1+e^{h^T w_j}}$

Problem 3 (33 points)

This problem considers a few different function classes that map from $\mathcal{X} = \Re^p$ to $\mathcal{Y} = \{0,1\}$.

(a) Consider the function class

$$\mathcal{H}_a = \{h : h(x) = [x == \mu]\,, \quad \mu \in \Re^p\}$$

where $[\cdot]$ is the unit indicator function. What is the VC dimension of class $\mathcal{H}_a$? Prove your answer: demonstrate that, with probability one, a training dataset $mathcalD = \{x_1, \ldots, x_n\}$ of size $n > d$ can be labeled in only $\mathcal{O}\left\{n^d\right\}$ different ways for VC dimension d.

SOLUTION: $d_{VC} = 1$. From any training dataset, it is possible to choose any one token to label as class 1; all others will be class 0. There are therefore at most $\mathcal{O}\{n\}$ ways to label a size-n dataset.

(b) Consider the function class

$$\mathcal{H}_b = \left\{h : h(x) = [\|x - \mu\| < b]\,, \quad \mu \in \Re^p, \quad b \in \Re^+\right\}$$

where $[\cdot]$ is the unit indicator function, and $\|x\|$ is the Euclidean norm. What is the VC dimension of class $\mathcal{H}_b$? Prove your answer: demonstrate that a training dataset $mathcalD = \{x_1, \ldots, x_n\}$ of size $n > d$ can be labeled in only $\mathcal{O}\left\{n^d\right\}$ different ways.

SOLUTION: $d_{VC} = (p + 1)$. This is a spherical classifier with threshold, which has the same VC dimension as a linear classifier with threshold. There are $\mathcal{O}\{n^p\}$ ways in which the dataset can be ordered, by choosing the centroid $\mu \in \Re^p$. Given a centroid, there are $\mathcal{O}\{n\}$ meaningfully different ways in which to choose the threshold. It can also be shown that this classifier is adequate to fracture every dataset of size $n = (p + 1)$, but is not adequate to fracture every dataset of size $n = p + 2$.

(c) Consider the function class

$$\mathcal{H}_c = \left\{h : h(x) = \left[\mu^T x \geq 1\right]\,, \quad \mu \in \Re^p\right\}$$

where $[\cdot]$ is the unit indicator function. What is the VC dimension of class $\mathcal{H}_c$? Prove your answer: demonstrate that a training dataset $mathcalD = \{x_1, \ldots, x_n\}$ of size $n > d$ can be labeled in only $\mathcal{O}\left\{n^d\right\}$ different ways, and/or demonstrate that a dataset can be labeled in 2^n different ways if and only if $n < d$.

SOLUTION: $d_{VC} = p$. This is a linear classifier without an explicit threshold. Although it seems to have the same flexibility as the linear classifier with threshold, it is relatively easy to show that this classifier can label any dataset with only $\frac{1}{2}$ as many different labelings as the linear classifier with explicit threshold. What is far less intuitively obvious (but true) is that the factor of two is sufficient to reduce the total number of labelings, for any dataset, to only $\mathcal{O}\{n^p\}$. One proof was given in the Vapnik-Chervonenkis paper; another was given in class. It is also possible to show that this classifier can fracture any dataset of size $n = p$, but that there are datasets of size $n = p + 1$ that cannot be fractured.