Machine Learning Midterm

This exam is open book. You may bring in your homework, class notes and textbooks to help you. You will have 1 hour and 15 minutes. Write all answers in the blue books provided. Please make sure YOUR NAME is on each of your blue books. Square brackets [] denote the points for a question. ANSWER ALL THREE QUESTIONS FOR FULL CREDIT

1. **Linear Algebra**

   (a) [7] Show for a matrix A and and eigenvector v that if $Av = \lambda v$ then

   $$A^k v = \lambda^k v$$

   Multiply both sides by $A$and then use the eigenvector equation. Repeat $k-1$ times.

   (b) [8] A colleague wants to know whether the dynamical system

   $$\dot{x} = Ax$$

   is stable where

   $$A = \begin{bmatrix} 3 & 1 \\ 2 & 2 \end{bmatrix}$$

   Show how to settle this question.
   Eigenvalue equation is

   $$(3-\lambda)(2-\lambda) - 2 = 0$$

   which has roots $\lambda = 1, 4$ both positive therefore unstable.

   (c) [10] In coding face images one can get the top M eigenvectors and eigenvalues. The same colleague now is thinking that some subset of these vectors might do a good job for a given two-class classification problem. What advice would you give him regarding the evaluation of the separation distance of the classes given a subset of the eigenvectors?

   One stratgy, use EM to fit two classes assuming Gaussian distributions. For different feature sets compare error rates.

   Another strategy, use a SUPPORT VECTOR MACHINE to estimate the separation distance for the two classes (assuming they separate).

2. **Information Theory**

(a) [5] A certain probability distribution for (x_1, x_2, x_3, x_4) is specified by $p(x_1) = \frac{1}{2}, (x_1) = \frac{1}{4}, (x_1) = \frac{1}{8}, (x_1) = \frac{1}{8}$. What is its entropy?

Entropy, H is given by

$$-\frac{1}{2}\log\frac{1}{2} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{8}\log\frac{1}{8} - \frac{1}{8}\log\frac{1}{8}$$

or

$$\frac{1}{2} + \frac{1}{4} \times 2 + \frac{1}{8} \times 3 + \frac{1}{8} \times 3 = \frac{7}{4}$$

(b) [5] Given a uniform distribution $q(x_i) = \frac{1}{4}, i = 1, \ldots, 4$, what is the Kullback-Liebler distance $KL_{p||q}$?

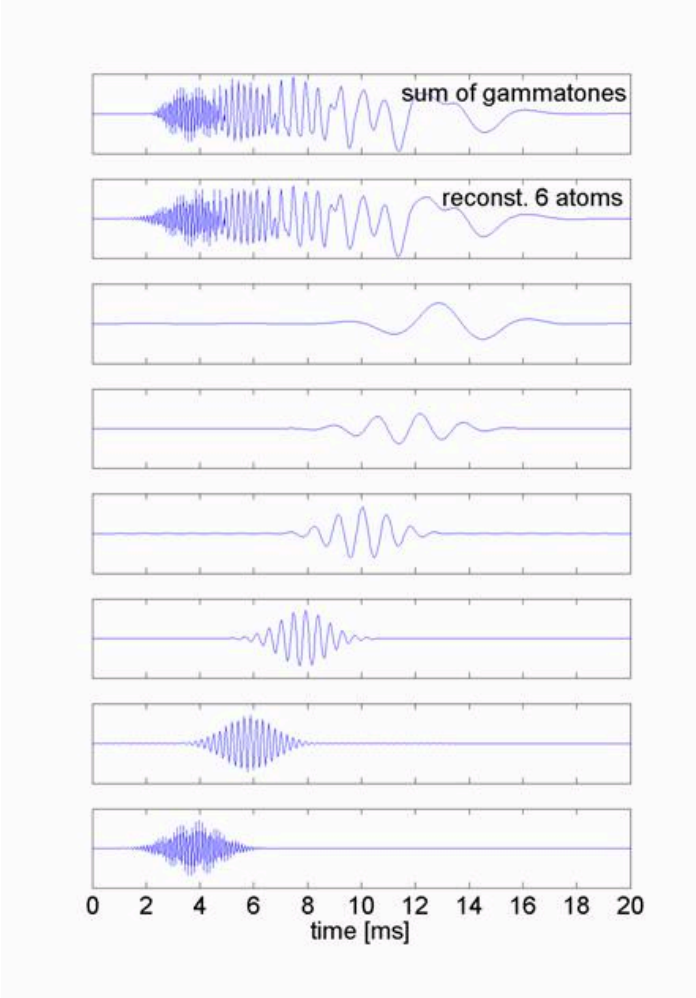
$$KL_{p||q} = \frac{1}{4}(-1 + 0 + 1 + 1) = \frac{1}{4}$$

(c) [5]Why is the KL distance useful?

Many algorithms searching for a good approximation p for a distribution can usefully compare it to a target distribution p.

(d) A given sound signal can be encoded by expressing the raw signal $x(t)$ as the sum of specialized functions called *gammatones* that are time-limited functions containing different frequencies. The figure below provides an example for $K = 6$ where the gammmatones in the **bottom six rows** can be used to code the small segment of sound source in **the top row**. More generally, a sound signal can be expressed as

$$x(t) = \sum_{k=1}^{K} \sum_{i=1}^{N_k} a_{ik}\, g_k(t - t_{ik})$$



i. [5]Suggest a way this code (or parts of it) could be used to recover mixed audio signals using the Independent Components (ICA) algorithm.

This not an easy question. Have to suggest some way of turning the data into a time varying scalar signal. could use the nearest gammatone index as the code. But the easiest way is to create N separate channels, one for each gammatone, apply ICA to each channel, and then idetify the recovered channels that go together by correlation and sum them.

ii. [5] Do you thinkyour answer to (i) would be an improvement over using the raw signals? Say why or why not.

The idea is that different sounds have different gammatine sequences. Since the gammatone capture frequencies better than the raw signal, then their frequency decomposition may work better.

3. **Optimization**

(a) [10] You are hired by a major soup company and they assign you the task of redesigning their cans. The goal is to maximize the volume V of the can for a given fixed surface area A. Luckily you remember that

$$V = \pi r^2 h$$

and

$$A = 2\pi r^2 + 2\pi r h$$

Solve this problem using the Lagrange Multiplier technique.

$$J = 2\pi r^2 h + \lambda(2\pi r^2 + 2\pi r h - A)$$

So that

$$J_r = 2\pi r h + \lambda(4\pi r + 2\pi h) = 0 \quad (1)$$

$$J_h = \pi r^2 + \lambda(2\pi r) = 0 \quad (2)$$

$$J_\lambda = 2\pi r^2 + 2\pi r h - A = 0 \quad (3)$$

From Eq. 2, $\lambda = -\frac{r}{2}$. Using this result in Eq. 1, $r = \frac{h}{2}$.
Using this result in Eq. 3,

$$r = \sqrt{\frac{A}{6\pi}}$$

(b) [5] Impressed with your can results, the company wants a consultation on a neural network problem. They have a network composed of 'neurons' with weights $\mathbf{w}$ such that the input $\mathbf{x}(k)$ is related to the output $y(k+1)$ where

$$y(k+1) = \sigma(\mathbf{w} \cdot \mathbf{x}(k))$$

i. [5] What is the advantage of the σ function compared to just using $y = \mathbf{w} \cdot \mathbf{x}(k)$?
The σ function introduces nonlinearity into the mapping process and as a result there are more ways of separating classes. The VC dimesion of such units is higher.

ii. [5] If you were to set up the Backpropagation algorithm for a network of sigmoid units as a classical optimization problem using the Lagrange Multipliers with a Hamiltonian, how many Lagrange multipliers would you need?
One for each "neuron."

iii. [5] Describe one of the potential pitfalls of the Backppropagation algorithm that your co-workers should look out for. How would SVMs deal with it?
One of:

- Slow to converge for networks with many layers. SVMs have only one layer
- No prescription on how to organize network. SVMs use Kernel functions that are easier to deal with.
- No tight bounds on the number of training samples. SVMs have smaller VC dimension values - have bounds.