CALD 10–702
Statistical Approaches to Learning and Discovery

Final Exam

May 2, 2003

Due Monday, May 12.

Instructions: This final exam/assignment includes four problems, involving decision theory, kernels, the Laplace approximation, and generative-discriminative models. You may use any reference materials while solving these problems, but please do not work in groups–all work must be your own. Pass in your written answers, including a printout of your code for the problems involving computation, by 5:00 p.m. on May 12, by bringing a hardcopy to Sharon Woodside in Wean 5315.

Problem 1. *Decision Theory*

In this problem you will explore some aspects of decision theory, in particular some basic relationships between admissibility and Bayes risk.

We begin by recalling some definitions and basic theory. The framework is decision making with a bounded loss function $0 \leq L(a, \theta) \leq k$, where $a \in \mathcal{A}$ is an action, and where $\theta \in \Theta$ is the parameter describing the unknown "state of nature." The decision maker's utility for action a, when the state of nature is θ, is $-L(a, \theta)$; thus, the decision maker is trying to minimize loss.

We assume that before choosing an action from $\mathcal{A}$, the decision maker may observe (without cost) the random quantity X whose conditional distribution given θ is $p(X \mid \theta)$. The *decision rule* $\delta \in \Delta$ specifies an action $\delta(X) \in \mathcal{A}$ for each value of X. Given θ, the *risk* of the decision rule $\delta(X)$ given the parameter is the conditional expected loss

$$E_\theta\left[L(\delta(X), \theta)\right] \;=\; \int_{\mathcal{X}} L(\delta(x), \theta)\, p(x \mid \theta)\, dx$$

With respect to a (proper) prior distribution $\pi(\theta)$, the *Bayes risk* $R(\pi, \delta)$ of the decision rule δ is the expected loss

$$R(\pi, \delta) \;=\; \int_{\Theta}\int_{\mathcal{X}} L(\delta(x), \theta)\, p(x \mid \theta)\, \pi(\theta)\, d\theta$$

A decision rule $\delta(X)$ is *Bayes for the prior* π if it minimizes the Bayes risk $R(\pi, \delta)$ with respect to $\delta \in \Delta$.

Finally, let $p(\theta \mid x)$ be the posterior distribution given that $X = x$ under the prior $\pi(\theta)$. The

Bayes posterior loss $R(\pi, \delta \mid x)$ of δ is is just the posterior mean of the decision rule's loss:

$$R(\pi, \delta \mid x) \;=\; \int_\Theta L(\delta(x), \theta)\, p(\theta \mid x)\; d\theta$$

1.1 Prove that a decision rule δ that minimizes the Bayes posterior loss for each x also minimizes the Bayes risk.

Next, recall that a decision rule δ is *admissible* if there is no other decision rule δ' such that

$$\begin{aligned} \text{for each } \theta \in \Theta, \quad E_\theta[L(\delta'(X), \theta)] \;&\le\; E_\theta[L(\delta(X), \theta)] \\ \text{for some } \theta \in \Theta, \quad E_\theta[L(\delta'(X), \theta)] \;&<\; E_\theta[L(\delta(X), \theta)] \end{aligned}$$

1.2 Assume that Θ is discrete and that the prior is non-zero, $\pi(\theta) > 0$, for each $\theta \in \Theta$. Prove that if δ minimizes the Bayes risk with respect to π then it is admissible.[1]

Problem 2. *Generative vs. Discriminative Models*

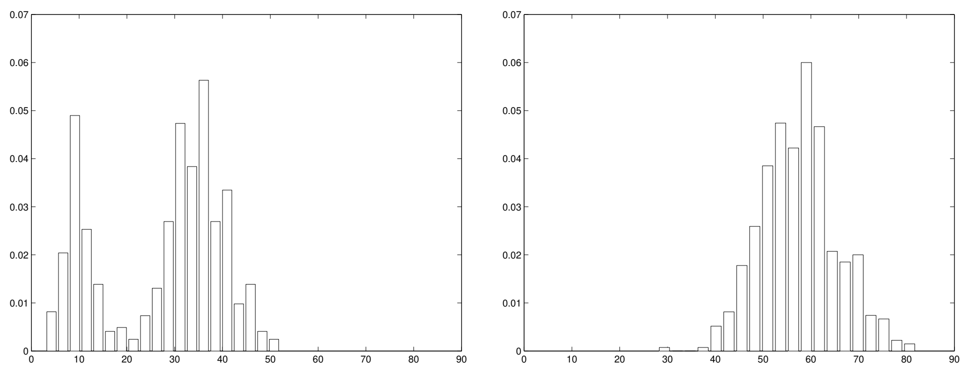
Suppose that we have a corpus of documents from two sources, Associate Press (AP) newswire data and transcriptions of National Public Radio (NPR) stories. We'd like to classify documents as to whether they're from AP or NPR *based only on their length.* A natural model of the length of a document is the Poisson: $p(k \mid \lambda) = e^{-\lambda}\frac{\lambda^k}{k!}$, for $k = 0, 1, 2, \ldots$.

2.1 *Generative model.* Explain in detail how this 2-class problem can be approached by modeling the lengths in each source with a Poisson distribution. Give the parameters, explain how the parameters are estimated, and how a document is classified once the models are fit.

2.2 *Discriminative model.* What is the discriminative model that corresponds to the generative model you derived in the previous problem? Give the discriminant function, the discriminative model, and describe how the parameters are estimated.

2.3 *Comparing the two.* Now implement your solutions above and test them on synthetic data that we have prepared. The file `prob2.class1.txt` in the directory `http://www.cs.cmu.edu/~tom/10-702/final/` contains samples from a Poisson mixture $p(k \mid \lambda_{1a}, \lambda_{1b}) = \lambda\, p(k \mid n, \lambda_{1a}) + (1 - \lambda)\, p(k \mid n, \lambda_{1b})$, and `prob2.class2.txt` contains samples from a single Poisson $p(k \mid \lambda_2)$. Here are histograms of the two data sets:

[1] The converse result is much more complicated, even in the case of a finite parameter space. See, e.g., (Berger, 1985, Theorem 5.8) or (Schervish, 1995, Theorem 3.95).



Divide up the data into random training and testing sets, with each training set containing a variable number of m points from each class. Train generative Poisson models and the corresponding discriminative models, and evaluate their error rates on the test set. Plot the results as error curves as a function of m, showing the standard errors. Is there a "crossover" point where the discriminative model outperforms the generative model? What conclusions do you draw about the relative merits of this generative model vs. the corresponding discriminative model on this data?

Problem 3. *Normal Approximations*

In a simple *genetic linkage model*,[2] animals are distributed into four categories: $Y = (y_1,\ y_2,\ y_3,\ y_4)$ according to the multinomial model given by $((2+\theta)/4,\ (1-\theta)/4,\ (1-\theta)/4,\ \theta/4)$.

3.1 What is the likelihood for the data $Y = (125, 18, 20, 34)$? What is the likelihood for the data $Y = (14, 0, 1, 5)$?

3.2 Use the Newton-Raphson algorithm to obtain the MLE $\widehat{\theta}$ of θ for $Y = (125,\ 18,\ 20,\ 34)$, and $Y = (14,\ 0,\ 1,\ 5)$. You may need to try different starting points to get convergence.

3.3 Plot the normalized likelihood and the associated normal approximation in the same figure for the data $Y = (125,\ 18,\ 20,\ 34)$. Repeat for $Y = (14,\ 0,\ 1,\ 5)$.

[2]From (Tanner, 1996, problem 2.9). Also, see Fisher's *Statistical Methods for Research Workers*, Section 52, for a simple explanation of the linkage model used in this problem. In short, the four multinomial categories give the expected counts (given θ) for two Mendelian traits in a self-fertilizing hybrid, each trait having a "3:1" dominant vs. recessive phenotype, under the possibility that the two traits are not independently distributed in the offspring. If the traits are independently distributed, then the expected counts in the four multinomial cells, corresponding to {doubly dominant, dominant in the first trait only, dominant in the second trait only, doubly recessive}, are in the ratio $9:3:3:1$. This corresponds to the independent product for two Mendelian traits under the "$3:1$" law for dominant vs. recessive phenotypes in a self-fertilizing hybrid, with $\theta = \frac{1}{4}$.

Problem 4. *Kernel Methods*

Suppose that we have a graph $G = (V, E)$ with vertex set V and edge set E; we'll let n denote the number of vertices. A *kernel* on G is simply a positive-definite $n \times n$ matrix $K = [K(i,j)]$ indexed by pairs of vertices in the graph, $i, j \in V$.

There is a natural family of graph kernels that we call *diffusion kernels*. Suppose that there is a weight on each edge in the graph, these weights yield a symmetric $n \times n$ matrix $W = [w_{ij}]$. Let D be the diagonal matrix with diagonal entries $D_{ii} = \sum_j w_{ij}$; this is the total weight on edges connected to node i. The *graph Laplacian* is defined as the matrix $\Delta = D - W$. The *diffusion kernel* K_β is then defined as

$$K_\beta = e^{-\beta\Delta}$$

where $e^{-\beta\Delta}$ indicates the matrix exponential.[3]

4.1 What is the feature space associated with the diffusion kernel K_β?

4.2 Describe a basis for the reproducing kernel Hilbert space defined by K_β.

Hint: For both of the problems above, think spectral. The matrix Δ can be shown to have only non-negative eigenvalues.

The standard Perceptron is the following algorithm:

PERCEPTRON

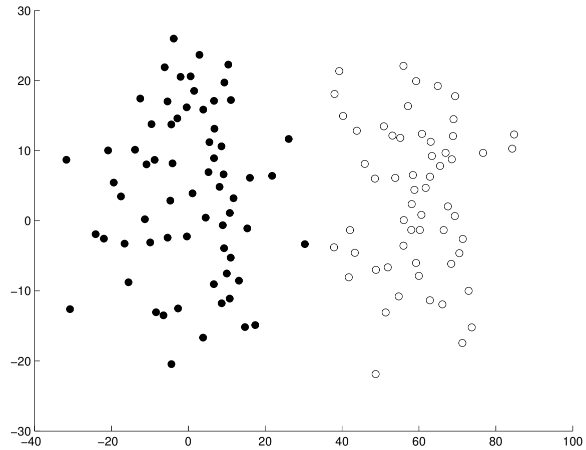
1. Input: $(x_1, y_1), \ldots, (x_n, y_n)$, with $x_i \in \mathbb{R}^m$, $y_i \in \{-1, +1\}$.
2. Initialize: $w_0 = 0$, $t = 0$
3. **While** $\exists\, i$ such that $y_i \neq \mathrm{sgn}(w_t^\top x_i)$:
4. Choose $i \in \{1, \ldots, n\}$ such that $y_i \neq \mathrm{sgn}(w_t^\top x_i)$
5. Update $w_{t+1} = w_t + y_i\, x_i$
6. $t = t + 1$
7. **end**
8. Output: w_t

Typically the data are simply processed sequentially.

4.3 Give the analogous perceptron algorithm when using a kernel, and show how the algorithm doesn't require the explicit use of feature space. (Derive the algorithm "from scratch" first, without looking it up or checking your notes!)

[3] I. Kondor and J. Lafferty, "Diffusion kernels on graphs and other discrete structures," ICML 2002, `http://www.cs.cmu.edu/~lafferty/ps/diffusion-kernels.ps`.

Your task in the remainder of the problem is to apply the perceptron algorithm to a synthetic data set, using a kernel. The directory `http://www.cs.cmu.edu/~tom/10-702/final/` contains data that you will apply the kernelized perceptron algorithm to. Here is what the data look like:



The file `prob4.mat` contains the (x, y) coordinates together with the class label for each of these 124 points, as well as a 124×124 weight matrix W for the complete graph on these points.

4.4 Use the kernel perceptron algorithm to learn the two-class problem given in the above data files. You may use the diffusion kernel $e^{-\beta\Delta}$ where $\Delta = D - W$, a polynomial kernel, or other kernel of your choice. (Better yet, try several kernels and compare the results.)

As for problem 2, divide up the data into random training and testing sets, with each training set containing a variable number of m points from each class. Train the kernel perceptron and evaluate its error rate on the test set. Plot the results as error curves as a function of m, showing the standard errors.