Name: Solution

**Massachusetts Institute of Technology**

# 6.891 Machine Learning

**Midterm** **Date: October 20, 1999**

---

| | |
|---|---|
| Problem 1 | 16 |
| Problem 2 | 10 |
| Problem 3 | 16 |
| Problem 4 | 12 |
| Problem 5 | 16 |
| Problem 6 | 16 |
| Total | 86 |

## Problem 1: Warmup

**A: (4 points)** TRUE or FALSE: The Perceptron Learning Algorithm converges for all learning problems:

False: The data must be linearly separable.

**B: (4 points)** Let

$$\sum_j |F(x^j, w) - y^j|$$

be a new learning criteria which penalizes the absolute difference between the multi-layer perceptron, $F(x^j, w)$ and the target $y^j$ (where $w$ is a vector of weights). Is backpropagation still applicable to the problem of estimating $w$? If so, what changes are necessary?

Backprop is still applicable. The only change is to the $\delta$s for the output units.

**C: (4 points)** TRUE or FALSE: Consider backpropagation learning of a multi-layer perceptron using batch updates and a sufficiently small learning rate. Will the algorithm always converge to the optimal weight vector for all learning problems? (The optimal weight vector yields the lowest error rate of all possible weight vectors.)

False: It will converge to a locally optimal solution.

## Problem 2: ML and MAP

**A: (4 points)** When is maximum a posteriori (MAP) estimation equivalent to maximum likelihood (ML) estimation and why?

MAP = ML when the prior is uniform over the parameters.

Alternatively MAP = ML when there is an infinite amount of data.

**B: (6 points)** Under what assumptions is finding the least squares fit of a function to data the same as finding the maximum likelihood function? I.E when is

$$\min_W (XW - Y)^T (XW - Y) \text{ equivalent to } \max_W P(Y|X, W)$$

$P(Y|X,W) = G(Y - W^T X)$

i.e. $Y$ is linearly related to $X$ plus Gaussian Noise.

$\left(\log G(Y - W^T X) \propto (W^T X - Y)^T (W^T X - Y)\right)$

## Problem 3: Risk

You are the President of General Motors and you must make a decision about the type of air bag system to place in new GM cars. The current system is very reliable and deploys the airbag in 100% of car accidents. Your research division has compiled statistics over 10,000 accidents regarding mortality rates for both adults and children:

| | With Airbags | Without Airbags |
|---|---|---|
| **Adults:** | | |
| Survive | 80% | 60% |
| Die | 20% | 40% |
| **Children:** | | |
| Survive | 75% | 80% |
| Die | 25% | 20% |

Your research division has also designed a new device which can sense whether the passenger in front of an airbag is a child or an adult. When a child is detected, the air bag is disabled. This device is correct regarding the age of the passenger 80% of the time. (Note: Adding this device costs no more than a standard air bag.)

**You must decide whether to include this new device in GM cars or not.**

**A: (4 points)** In order to make this decision several pieces of information are missing. What information is missing? (Note: one could argue that all sorts of information is missing: whether the world will come to an end on 1/1/2000 for example. We are looking for the minimal amount of information that is missing so that this decision can be made and justified.) Choose concrete numerical values for this information (based on your intuitions).

$P(Adult) = 1 - P(child) = 0.5$

Loss: Child Dies: 2
Child Lives: 0
Adult Dies: 1
Adult Lives: 0

P(Accident) is interesting, but it doesn't change your answer.

**B: (6 points)** Based on the information above and your numerical assumptions, would you install these new devices in cars? The alternative is to install airbags without these new child sensing devices.

Expected loss without Device

$2 \times P(\text{Child Dies} \mid \text{Accident}) + 1 \times P(\text{Adult Dies} \mid \text{Acc})$

$2 \times 0.5 \times 0.25 + 1 \times 0.5 \times 0.20 = 0.35$

With Device

$$2 \times \left[ \begin{matrix} P(\text{Ch Dies \& Detected}) \\ + P(\text{Ch Dies \& Not Det}) \end{matrix} \right] + 1 \times \left[ \begin{matrix} P(\text{Adult Dies \& Det}) \\ + P(\text{A Dies \& } \overline{\text{Det}}) \end{matrix} \right]$$

$$2 \times \left[ \begin{matrix} 0.8 \times .20 \\ + 0.2 \times .25 \end{matrix} \right] \times 0.5 + 1 \times \left[ \begin{matrix} 0.8 \times 0.2 \\ + 0.2 \times 0.4 \end{matrix} \right] \times 0.5$$

$= 0.21 + 0.12 = 0.33$

Yes

**C: (6 points)** Based on the information above and your numerical assumptions, what is the minimum acceptable accuracy for the child sensing device? (i.e. the smallest value for accuracy which would make you decide to install the device.)

find pt where risks equal

$$0.35 = 2\left[ \begin{matrix} \lambda(0.2) \\ +(1-\lambda)0.25 \end{matrix} \right] 0.5 + 1\left[ \begin{matrix} \lambda(0.2) \\ +(1-\lambda)0.4 \end{matrix} \right] 0.5$$

$$= 0.25 - \lambda(0.05) + 0.2 - \lambda(0.1)$$

$$= 0.45 - \lambda(0.15)$$

$$\Rightarrow \lambda \geq \frac{0.10}{0.15} \approx 67\%$$

.16
.125
.08
.04
.385

.16
.125
.285

.016
.050
.066

$2 \times 0.5 \times 0.66 = 0.066$

$1 \times 0.5 \times .24 + 0.12$

.16
.08
.24

.786

## Problem 4: Exponential Distributions?

You are given a new type of density which has two parameters $\mu$ and $s$:

$$p_1(x) = ce^{-[\frac{x-\mu}{s}]^4} \tag{1}$$

**A: (4 points)** Write an expression for $c$'s value. Does $c$ depend on $\mu$? What is the qualitative difference between $p_1(X)$ and a Gaussian?

$\int_{-\infty}^{\infty} P_1(x)\,dx = C$ No it does not depend on $\mu$, because the limits go from $-\infty$ to $+\infty$

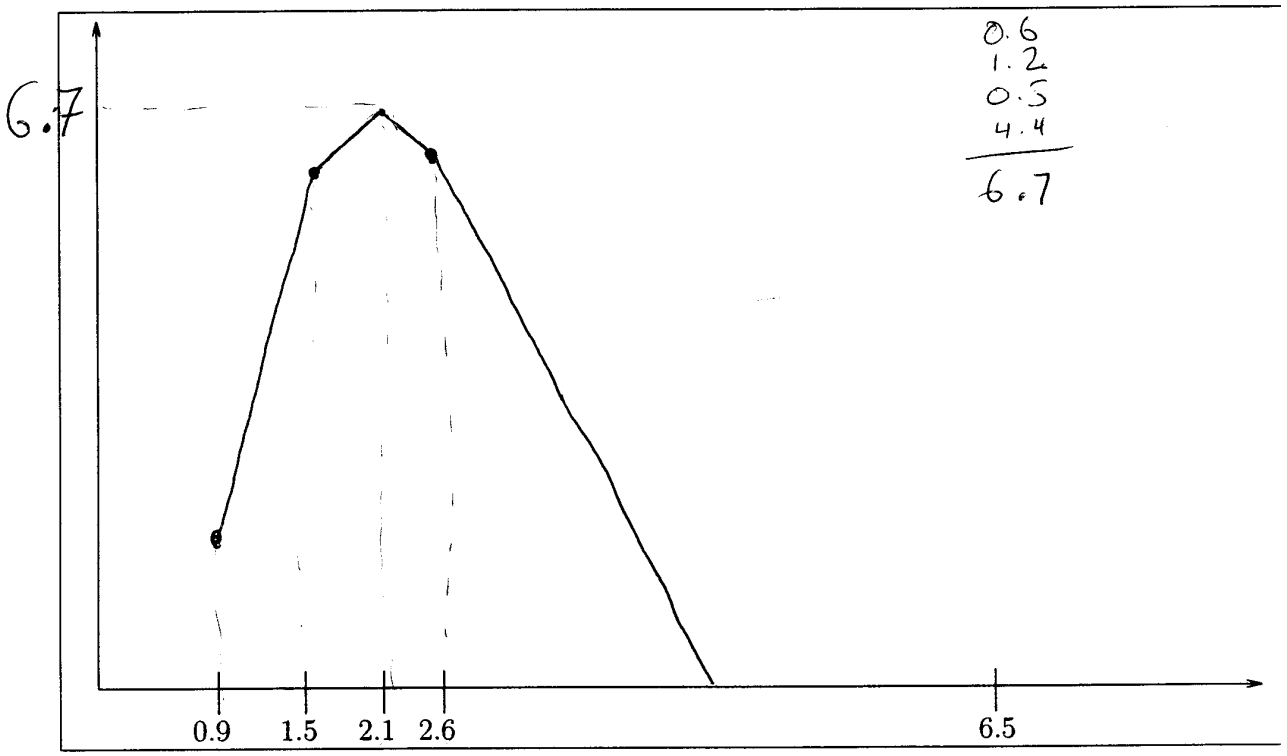
$P_1(x)$ falls off to zero more quickly than a Gaussian.

Consider the density:

$$p_2(x) = ke^{-|x-\mu|} \tag{2}$$

where $|x - \mu|$ is the absolute value of $|x - \mu|$.

**B: (4 points)** You are given the examples 0.9, 1.5, 2.1, 2.6, 6.5. On the graph below draw a curve which shows the log likelihood of the data given $\mu$. While neatness does count, what is more important is that you get the qualitative form of the graph correct. Please include some comments in case we cannot figure out what you mean from the graph.




**C: (4 points)** What is the maximum likelihood estimate for $\mu$?

2.1 (the median)

## Problem 5: ML and Mixtures of Gaussians

**A: (4 points)** Given 4 1D points 0.8, 1.2, 4.0, 5.0, 5.8, fit them with a 2 component mixture of Gaussians density. Use the maximum likelihood criteria. (If you feel that there is more than one equivalent model, indicate this, but give us at least one concrete answer.)

$\mu_1 = 0.8$ $\mu_2 = 4.0$

$\sigma_1 = 0.0$ $\sigma_2 = 3.02$

1.2
4.0
5.0
+ 5.8
16.0

**B: (6 points)** True or False: now suppose you also know that $p(\mu_1)$ and $p(\mu_2)$ are uniform densities over the range $[0, 10]$. and that $p(\sigma_1)$ and $p(\sigma_1)$ uniform densities over the range $[0.5, 10]$. Would this change the answer you gave above significantly? (using the maximum a posteriori criteria). Give one or two sentences of justification.

Yes. $\sigma_1 = 0.0$ is no longer allowed
Therefore the 2 gaussians must fit the data more equally.

**C: (4 points)** True or False: We saw that EM can be used to find the maximum likelihood mixture of Gaussians through a sequence of steps which can be implemented efficiently. EM can also use used to estimate other type of mixture models as well. Unfortunately application of EM to some mixture models does not always lead to efficient computations. Give one reason why this might be so.

M step maxizes the likelihood of the data. This might not be a straightforward calculation (non-exponential densities in the mixture model)

**D: (6 points)** Your buddy Fred proposes that rather than the MAP model it is always better to use a Bayesian approach for density estimates. Write down a complete expression for this estimate using all the information given above (which includes the data, the priors, and the form of the density estimator). You can write this as a single expression, but it is probably easier to define sub-expressions as well.

$$P(x \mid \mathcal{X}) = \int P(x \mid \theta)\, P(\theta \mid \mathcal{X})\, d\theta \qquad \mathcal{X} = \text{data},\ \theta = \text{params}$$

$$P(x \mid \theta) = \tfrac{1}{2}\left[G(x, \mu_1, \sigma_1) + G(x, \mu_2, \sigma_2)\right]$$

$$P(\theta \mid \mathcal{X}) = \frac{P(\mathcal{X} \mid \theta)\, P(\theta)}{P(\mathcal{X})}$$

$$P(\mathcal{X} \mid \theta) = \prod_i P(X = x^i \mid \theta)$$

$$P(\mathcal{X}) = \int P(\mathcal{X} \mid \theta)\, P(\theta)\, d\theta$$

## Problem 6: Discriminants

**A: (4 points)** For a two class discrimination problem where each class is a gaussian the resulting decision boundary can sometimes be expressed as a linear discriminant (e.g. $\theta(w^T x + b)$). What conditions must be satisfied for this to be the case?

$\Sigma_1 = \Sigma_2$: equal covariances.

In class we saw two criteria for learning linear discriminants directly: minimum squared error and the Perceptron criteria. Will either of these criteria yield the same discriminant function as you would have observed if you had first estimated two gaussian densities and then used Bayes' rule? Please be as formal and direct as possible. In one of these cases the formal proof is quite compact... very little discussion is necessary.

**B: (6 points)** Is the gaussian discriminant the same as the minimum squared error discriminant?

Yes, but the derivation is involved.

→ The book proves MSE is equivalent to Fisher's Linear discriminant.

Fisher's is equivalent to 2 Gaussians.

A direct proof is related to,

**C: (6 points)** Is the gaussian discriminant the same as the Perceptron discriminant?

(1) Perceptron only works for separable data.

(2) There are many potential Perceptron discriminants (for linearly separable data). There is only 1 gaussian discriminant.