EE660 Nov 28, 2018

Discussion 15

Today:

- Reminder: course evaluations
- Project report template
- Review exercises

Problem 4

$$p(y|\underline{x},\underline{w}) = \frac{y}{b^2} e^{-y/b}, \quad y>0, \quad b = \underline{w}^T\underline{x} > 0$$

a) $NLL = -\ln p(D|\underline{\theta})$

$$= -\ln \prod_{i=1}^{N} p(y_i|x_i, w)$$

$$= -\sum_{i=1}^{N} \ln\left[\frac{y_i}{b_i^2} e^{-y_i/b_i}\right] =$$

$$= \sum_{i=1}^{N}\left[-\ln y_i + 2\ln \underline{w}^T\underline{x}_i + \frac{y_i}{\underline{w}^T\underline{x}_i}\right]$$

Problems 1-3, and Problem 6, are short-answer problems.

Problem 1 (10 points)

Consider a regression problem that uses MAP estimation to find the weights $\underline{w}$ and if applicable also the binary mask vector $\underline{\gamma}$. The model is:

$$p\left(y \middle| \underline{x}, \underline{\theta}\right) = N\left(y \middle| \underline{w}^T \underline{x} + w_0, \sigma^2\right)$$

with σ^2 given. For each prior given in parts (a)-(e) below, find the matching regularizer term(s) from (i)-(vi) below. Note that in cases that include the parameter $\underline{\gamma}$, in the "regularizer terms" column below (and in the model above) we redefine $\underline{w}$ as:

$$\left(\underline{w}_\gamma\right)_j = \gamma_j w_j, \quad \gamma_j \in \{0,1\}$$

and then drop the γ subscript so that $\underline{w} \leftarrow \underline{w}_\gamma$.

Hint: do not assume that all regularizer terms will be used in this problem.

Parts (a) - (e) are worth 2 points each.

Priors:

(a) $p(\underline{w}) = \prod_{j=1}^{D} N\left(w_j \middle| 0, \sigma_w^2\right)$

Answer: v

(b) $p(\underline{w}) = \prod_{j=1}^{D} \text{Lap}\left(w_j \middle| 0, \frac{1}{\lambda'}\right) \propto \prod_{j=1}^{D} e^{-\lambda' |w_j|}$

Answer: iv

(c) $p(\underline{w}, \underline{\gamma}) = N\left(\underline{w} \middle| \underline{0}, \sigma_w^2 \underline{\underline{I}}\right) \pi_0^{\,q} \left(1 - \pi_o\right)^{(D-q)}$

in which $q = \sum_{j=1}^{D} \gamma_j$, and $0 \le \pi_0 \le 1$.

Answer: i

Regularizer terms:

(i) $\lambda \|\underline{w}\|_0$

(ii) $\lambda \sum_{j=1}^{D} |w_j|^{1/3}$

(iii) $\lambda \sum_{j=1}^{D} |w_j|^{1/2}$

(iv) $\lambda \|\underline{w}\|_1$

(v) $\lambda \|\underline{w}\|_2^2$

(vi) $\lambda \sum_{j=1}^{D} |w_j|^3$

Problem 2 continues on next page...

(d) $p(\underline{w}) = \prod_{j=1}^{D} \frac{1}{4\Gamma(4)} e^{-|w_j|^{1/3}}$

Answer: ii

(e) $p(\underline{w}) = \prod_{j=1}^{D} \frac{1}{\Gamma\left(\frac{3}{2}\right)} e^{-|w_j|^{2}}$

Answer: ✓

Problem 2 (12 points)

You are solving a 2-class classification problem using random forests. There are 2 features, and feature space is $0 \le x_1 \le 10$, $0 \le x_2 \le 10$. Each tree is found by randomly picking one of the features, and only one iteration is performed (that is, each resulting tree has 1 root and two leaf nodes). You use a total of 3 trees. The resulting decision regions of the 3 trees are:

$$T_1:\ 0 \le x_1 \le 3 \Rightarrow \hat{y}^{(1)} = 1;\ \ 3 < x_1 \le 10 \Rightarrow \hat{y}^{(1)} = 2$$

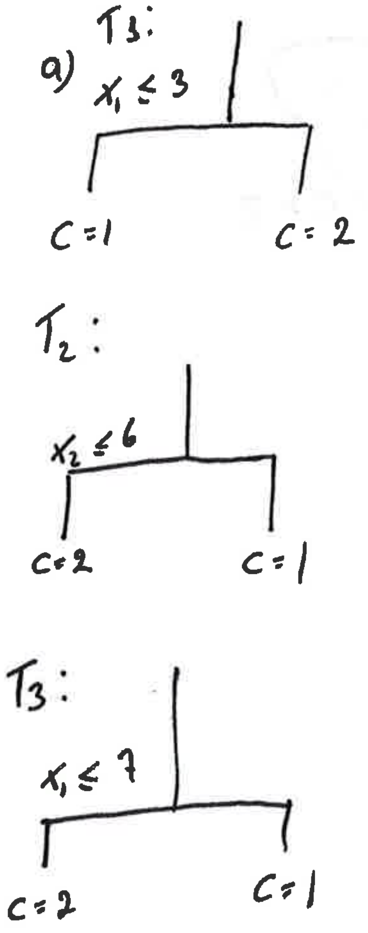
$$T_2:\ 0 \le x_2 \le 6 \Rightarrow \hat{y}^{(2)} = 2;\ \ 6 < x_2 \le 10 \Rightarrow \hat{y}^{(2)} = 1$$

$$T_3:\ 0 \le x_1 \le 7 \Rightarrow \hat{y}^{(3)} = 2;\ \ 7 < x_1 \le 10 \Rightarrow \hat{y}^{(3)} = 1$$

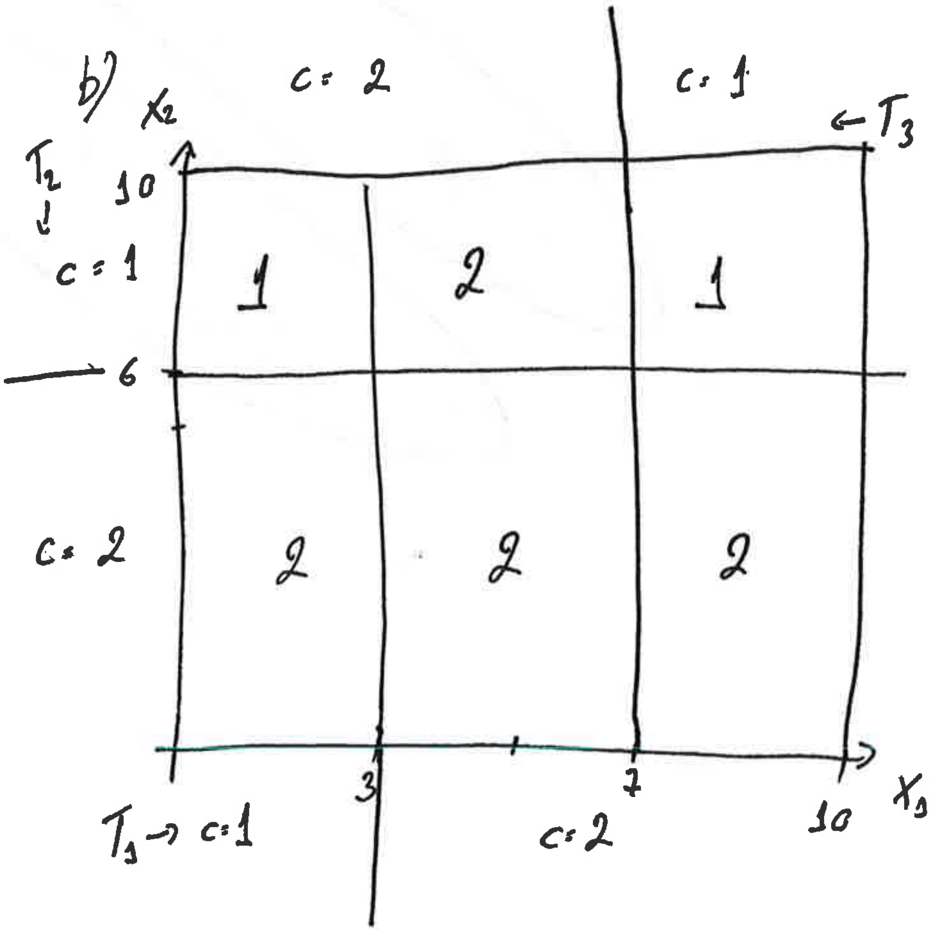
(a) (3 points) Draw the three decision trees.

(b) (9 points) In 2D feature space, draw the final decision boundaries and label the final decision regions that result from the random forest algorithm. Optionally, briefly describe your approach for partial credit.

a)



b)



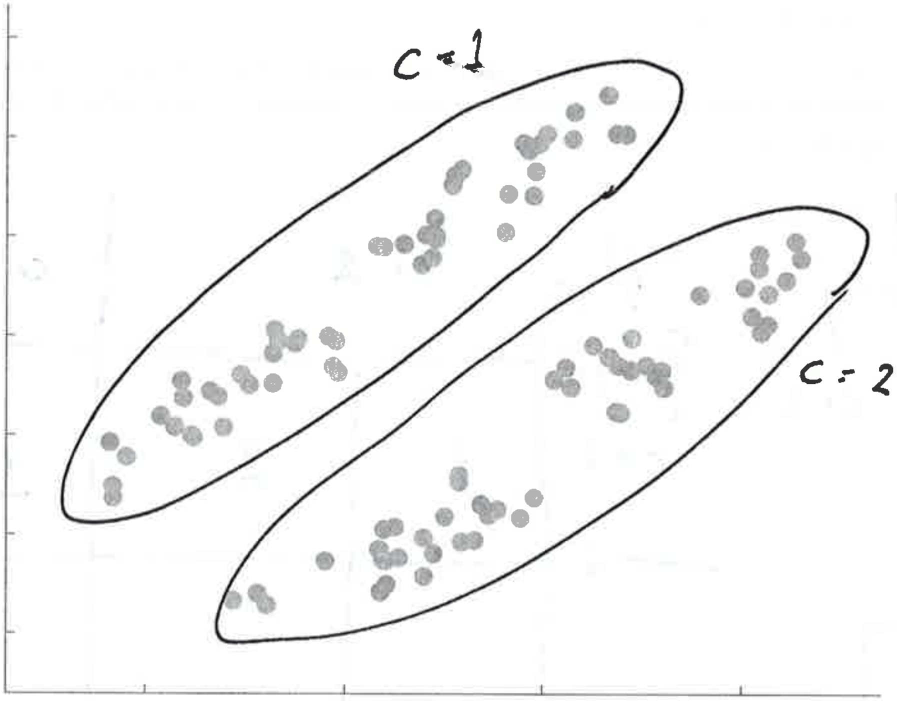
5 4

Problem 3 (10 points)

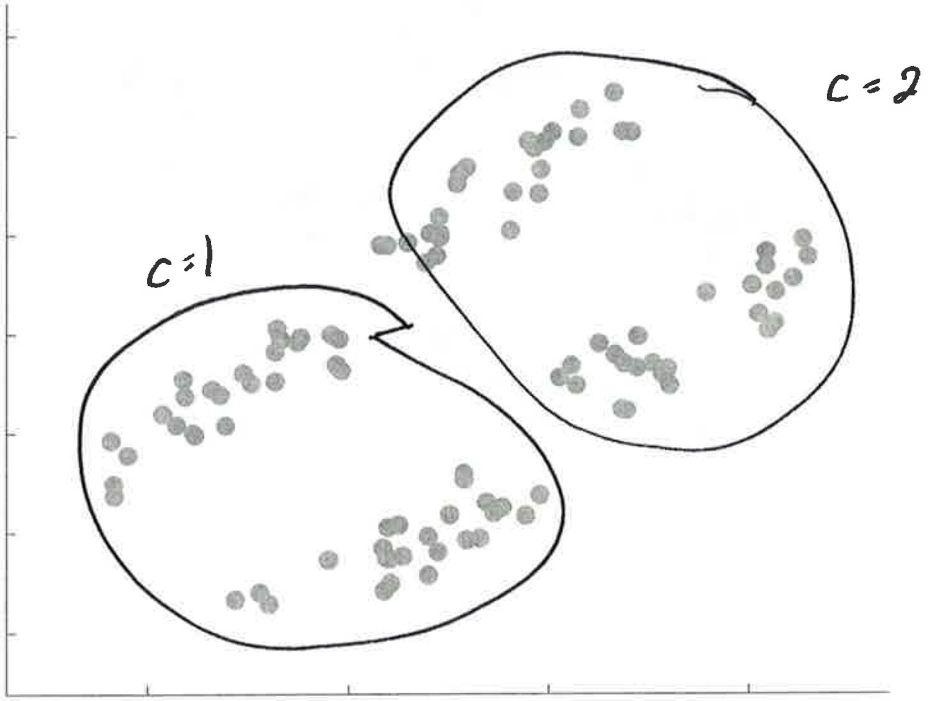
You are given a set of data points for a 2D unsupervised learning problem, plotted below. You will consider two hierarchical graphical clustering algorithms applied to this data.

Note that all plots given below show identical data. Also, there is an extra page after this problem with two more plots, in case you need them. **Be sure to indicate clearly which plots show your final answers, and for which part.**

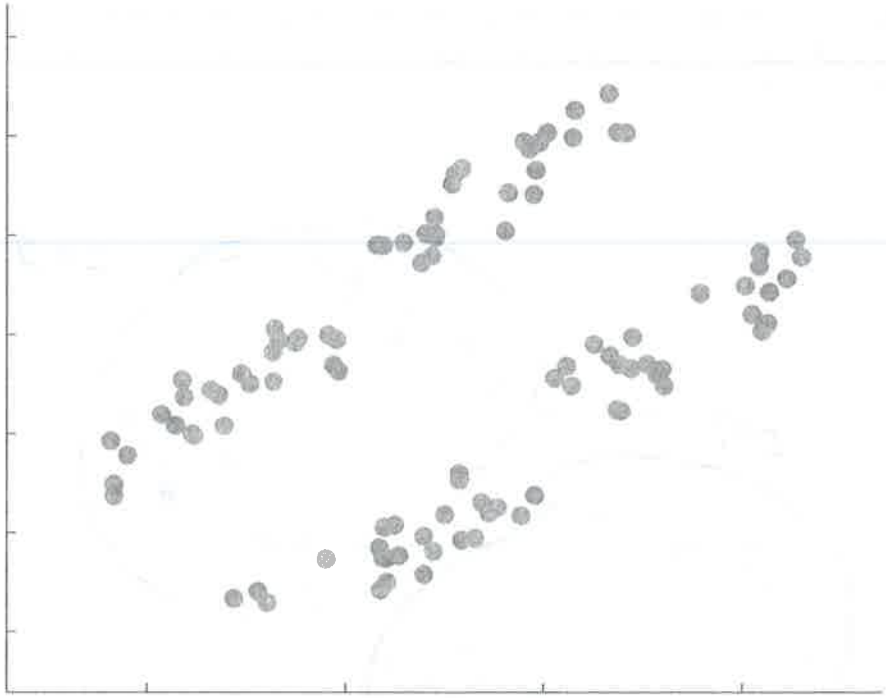
(a) (5 points) For the nearest-neighbor (single linkage) algorithm taken to $K = 2$ clusters, based on your understanding of the algorithm, predict what you think the resulting clusters would be. You may label them by enclosing each cluster with one or more closed curves, and labeling the interior of each curve as $c = 1$ or $c = 2$. Note: no need to actually apply the algorithm to the data.

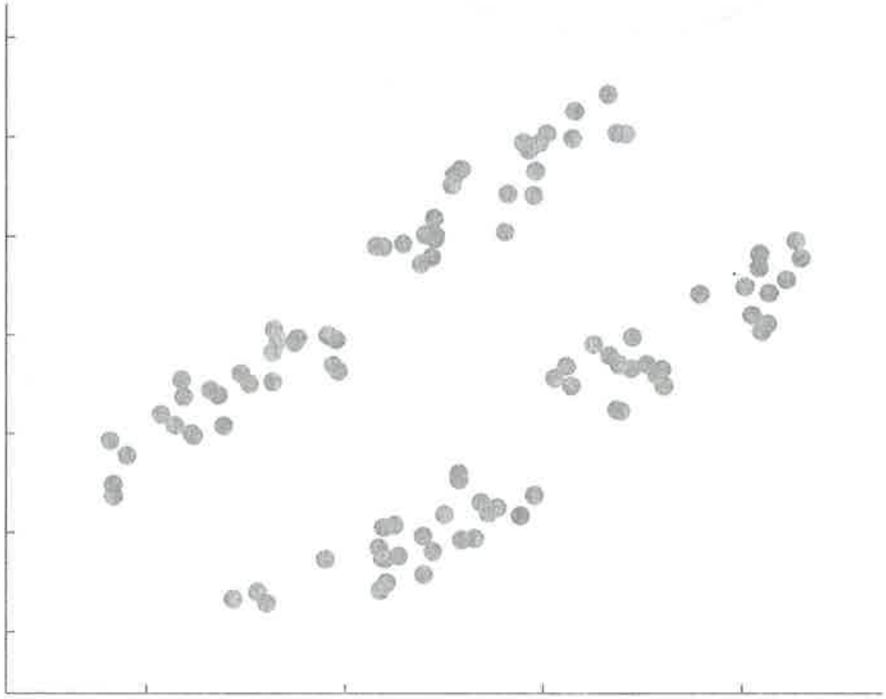


(b) (5 points) For the farthest-neighbor (complete linkage) algorithm taken to $K = 2$ clusters, based on your understanding of the algorithm, predict what you think the resulting clusters would be. You may label them by enclosing each cluster with one or more closed curves, and labeling the interior of each curve as $c = 1$ or $c = 2$. Note: no need to actually apply the algorithm to the data.



Extra plots for Problem 2 (Indicate part (a) or part (b) for each.)





Problem 4 (24 points)

You want to use regression with MAP estimation of the weights to solve a problem. You have enough information about the problem to know the output y is always positive, and would be better modeled as a gamma distribution (with particular choice of parameters given below) than a Gaussian distribution; thus:

$$p\left(y \middle| \underline{x}, \underline{w}\right) = \frac{y}{b^2} e^{-y/b}, \quad y > 0, \quad b = \underline{w}^T \underline{x} > 0 .$$

All the feature values are also known a priori to be positive, so by requiring $\underline{w}_j > 0 \ \forall j$ we can ensure the constraint on b is satisfied.

You have a training dataset $\mathcal{D} = \left\{\left(\underline{x}_i, y_i\right)\right\}_{i=1}^{N}$.

(a) (6 points) Write the negative log likelihood in terms of given parameters. Simplify as much as possible. Your answer should be in terms of $\underline{x}_i$, y_i, and $\underline{w}$.

(b) (10 points) We choose a prior given by:

$$p\left(\underline{w}\right) = \prod_{j=1}^{D} p\left(w_j\right) = \prod_{j=1}^{D} \left\{ r e^{-r w_j} \left[\!\left[w_j > 0 \right]\!\right] \right\}, \qquad r > 0$$

in which $\left[\!\left[\bullet \right]\!\right]$ denotes the indicator function.

Write the objective function f_{obj} in terms of given quantities. Set up f_{obj} so that its **minimum** will yield the MAP solution. Simplify your expression as much as possible, keeping in mind that the minimization will be taken w.r.t. $\underline{w}$. **No need** to carry out the minimization to find the optimal $\underline{w}$. **Hint:** if you have a term of the form $\ln\left[\!\left[w_j > 0 \right]\!\right]$, you can leave it in that form in your final answer.

(c) (4 points) Does the objective function have a regularizer term? If so, state what the regularizer term is, and answer: will it tend to reduce values of the weights, increase values of the weights, or neither?

(d) (4 points) If your objective function has any terms with $\left[\!\left[w_j > 0 \right]\!\right]$ or similar function of w_j or $\underline{w}$ in them, describe the effect of such terms on the minimization of f_{obj} and on the resulting optimal value of $\underline{w}$.

Problem 5 (20 points)

In this problem you will consider boosting applied to a 2-class classification problem. This problem is based on the general Forward Stagewise Additive Modeling algorithm framework, discussed in class, and using an exponential loss. For your convenience, this algorithm framework is given (in algebraic form) on the next page.

In this problem each simple classifier $\phi(\underline{x};\underline{\gamma}_m)$ is a decision stump - a 1-stage CART that applies a threshold value to one feature.

(a) (2 points) Define the exponential loss function $L_{\exp}\left(y_i, f(\underline{x}_i)\right)$ in terms of its arguments.

(b) (2 points) Write step 2(i) of the Forward Stagewise Additive Modeling algorithm in the form:

$$\underset{\beta_m,\underline{\gamma}_m}{\arg\min}\sum_{i=1}^{N} w_{im}\exp\left[-y_i\beta_m\phi(\underline{x}_i;\underline{\gamma}_m)\right],\quad \beta_m>0,\ \phi\in\{-1,+1\},\ y_i\in\{-1,+1\}$$

and define w_{im} .

(c) (12 points) At the m^{th} iteration, assume that $\underline{\gamma}_m$ (and therefore $\phi(\underline{x}_i;\underline{\gamma}_m)$) is given, and derive the optimal β_m by carrying out the minimization w.r.t. β_m . Use the following definitions to simplify the notation:

$$S_C=\sum_{\text{all } i\in CC} w_{im}, \qquad S_I=\sum_{\text{all } i\in IC} w_{im}$$

in which CC denotes the set of indexes i of data points $\underline{x}_i$ that are *correctly* classified by $\phi(\underline{x}_i;\underline{\gamma}_m)$, and IC denotes the set of indexes i of data points $\underline{x}_i$ that are *incorrectly* classified by $\phi(\underline{x}_i;\underline{\gamma}_m)$. Give the resulting optimal β_m in simplified form, expressing your answer in terms of S_C and S_I . Show your work.

(d) (4 points) From your answer to part (c), what does $\beta_m>0$ imply?

b) $p(\underline{w}) = \prod_{j=1}^{D} p(w_j) = \prod_{j=1}^{D} \{ r e^{-rw_j} [\![w_j > 0]\!] \}$, $r > 0$

$\downarrow$

indicator function

$$f_{obj}(w) = -\ln[p(D|\underline{\theta}) . p(\underline{\theta})]$$

$$= -\ln p(D|\underline{\theta}) - \ln p(\underline{\theta}) \quad , \quad \underline{\theta} = \underline{w}$$

$$\ln p(w) = \sum_{j=1}^{D} \ln r - rw_j + \ln[\![w_j > 0]\!]$$

$$f_{obj}(w) = \sum_{i=1}^{N} \left[2\ln(w^T x_i) + \frac{y_i}{\underline{w^T x_i}} \right] + r \sum_{j=1}^{D} w_j - \sum_{j=1}^{D} \ln[\![w_j > 0]\!]$$

c) Yes, $r \sum_{j=1}^{D} w_j$ $(= r\|\underline{w}\|_1$ because $w_j > 0 \ \forall j)$

Tends to reduce w_j's.

d) $- \sum_j [\![w_j > 0]\!] = +\infty$ for any $w_j \leq 0$, so will ensure all $w_j > 0$.

If $w_j > 0$, the term has no effect.