ECE523: Engineering Applications of Machine Learning and Data Analytics
Due 01/26/2018 @ 11:59PM (D2L)

I acknowledge that this homework is solely my effort. I have done this work by myself. I have not consulted with others about in any way. I have not received outside aid (outside of my own brain). I understand that violation of these rules contradicts the class policy on academic integrity.

Name: David Akre

Signature: David Akre

Date: 1/24/18

Instructions: There are seven problems. Partial credit is given for answers that are partially correct. No credit is given for answers that are wrong or illegible. All work must be supported and code must be submitted for credit.

Theory: ________

Practice: ________

Total: ________

Machine Learning
HW #1

David Akre
ECE 523
1/26/18

① Maximum Posterior vs. Probability of choice

(a) Show/Explain that $p(\omega_{max}|x) \geq \frac{1}{c}$ when we are using the Bayes Decision Rule. Derive an expression for $p(error)$. Let ω_{max} be the state of nature for which $p(\omega_{max}|x) \geq p(\omega_i|x)$ for $i = 0, 1, \dots, c$. Show that $p(error) \leq (c-1)/c$ when we use the Bayes decision rule to make a decision

Step 1: Derive an expression for $p(error)$

$$p(error) = \int_1^c p(error, x)\,dx \quad \leftarrow \therefore p(error, x) = p(error|x)\,p(x)$$

$$= \int_1^c p(error|x)\,p(x)\,dx \quad \leftarrow \text{Choose class with } \underline{\underline{\text{highest posterior}}} \text{ (i.e. } 1 - p(\omega_{max}|x))$$

$$= \int_1^c (1 - p(\omega_{max}|x))\,dx$$

$$= \int_1^c dx - \int_1^c p(\omega_{max}|x)\,dx$$

$$= x\Big|_1^c - p(\omega_{max}|x)\,x\Big|_1^c$$

$$= x(1 - p(\omega_{max}|x))\Big|_1^c$$

$$= c(1 - p(\omega_{max}|x)) - (1 - p(\omega_{max}|x))$$

$$= c - c\,p(\omega_{max}|x) - 1 + p(\omega_{max}|x) \quad \leftarrow \therefore \text{Divide by } c \text{ and cancel } p(\omega_{max}|x)$$

$$= \frac{c-1}{c}$$

$$\boxed{\therefore p(error) \leq \frac{c-1}{c}}$$

Step 2: Show that $p(\omega_{max}|x) \geq \frac{1}{c}$ for Bayes Decision Rule

$$p(\omega_{max}|x) = \frac{p(x|\omega_{max})\,P(\omega_{max})}{P(x)}$$

$\leftarrow$ $p(x|\omega_{max})$ is the likelihood probability, $P(\omega_{max})$ is the prior probability, and $p(x)$ is the conditional or evidence

$\therefore$ ω_{max} resembles the class with the highest posterior probability which in turn means that it resembles the minimum probability of error. So for $i = 0, 1, \dots, c$ $p(\omega_{max}|x) \geq p(\omega_i|x)$ means that $\underline{p(error) \leq \frac{c-1}{c}}$.

$$1 - p(error|x) \geq p(\omega_i|x)$$

$$1 \geq P(error|x) + p(\omega_i|x) \quad \leftarrow \text{NOTE: The error summed with } p(\omega_i|x) = 1$$

$$1 \geq p(error|x) + \sum_{i=1}^{c} p(\omega_i|x) \quad \leftarrow \lim \sum_{i=0}^{c} p(\omega_i|x) = 1 - p(\omega_{max}|x) \text{ or } \frac{1}{c}$$

NOTE: The posterior probability must be greater than or equal to the inverse of the number of classes, otherwise a different class is the real ω_{max} (e.g. $\sum_{i=0}^{c} p(\omega_i|x) \leq p(\omega_{max}|x)$)

$$1 \geq 1 - p(\omega_{max}|x) + \frac{1}{c}$$

$$-\frac{1}{c} \geq -p(\omega_{max}|x)$$

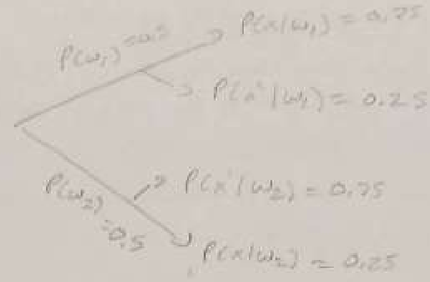
$$\boxed{\frac{1}{c} \leq p(\omega_{max}|x)}$$

Step 3: show proof in an example

(ex)(i) - Two classes : ω_1 & ω_2

- $P(\omega_1) = 0.5$
- $P(\omega_2) = 0.5$
- $P(x|\omega_1) = 0.75$
- $P(x|\omega_2) = 0.25$

(ii) (ex) Decision Tree



(iii) Total probability (evidence) $p(x) = p(\omega_1)p(x|\omega_1) + p(\omega_2)p(x|\omega_2)$

$= (0.5)(0.75) + (0.5)(0.25)$

$= \underline{\underline{0.5}} \rightarrow \underline{\underline{p(x') = 0.5}}$

(iv) Distribution for all events represented by the following table:

$p(x|\omega_1)\,p(\omega_1) = 0.375$
$p(x'|\omega_1)\,p(\omega_1) = 0.125$
$p(x|\omega_2)\,p(\omega_2) = 0.125$
$p(x'|\omega_2)\,p(\omega_2) = 0.375$

(v) Using Bayes theorem find $P(\omega_1|x)$ & $p(\omega_2|x)$:

$$p(\omega_1|x) = \frac{p(x|\omega_1)P(\omega_1)}{P(x)} = \underline{\underline{0.75}}$$

$$p(\omega_2|x) = \frac{p(x|\omega_2)P(\omega_2)}{P(x)} = \underline{\underline{0.25}}$$

(vi) Find ω^* using Bayes Decision Rule:

— $P(\omega_1|x) \geq p(\omega_2|x)$; so $\underline{\underline{p(\omega_1|x)}}$ is selected since it has a higher posterior probability

$\therefore$ Check $p(\omega_1|x) \geq \frac{1}{c}$ where $c = 2$

$\underline{\underline{0.75 \geq 0.5}}$ ← True (first proof holds)

(vii) Verify $p(error) \leq \frac{c}{c-1}$

— $p(error|x) = 1 - p(\omega max|x)$; $\omega_{max} = \omega_1$ — $p(error|x) = 1 - p(\omega_2|x)$

$= 1 - 0.75$ $\qquad$ $= 1 - 0.25$

$= \underline{\underline{0.25}}$ $\qquad$ $= \underline{\underline{0.75}}$

$\therefore$ $0.25 < 0.75$ (ω_1 leads to a smaller probability of error)

— $0.25 \leq \frac{2-1}{2}$

$0.25 \leq 0.5$ ← True (second proof holds)

2 Bayes Decision Rule Classifier

(a) Let the elements of a vector $x = [x_1 \dots x_d]^T$ be binary valued. Let $P(\omega_j)$ be the prior probability of the class ω_j ($j \in [1,2]$) and let $p_{ij} = P(x_i = 1 | \omega_j)$ with all elements in x being independent. If $P(\omega_1) = P(\omega_2) = \frac{1}{2}$, and $p_{i1} = p > \frac{1}{2}$ and $p_{i2} = 1-p$, show that the minimum error decision rule is: choose ω_1 if $\sum_{i=1}^{d} x_i > \frac{d}{2}$. HINT: choose ω_1 if $P(\omega_1)P(x|\omega_1) > P(\omega_2)P(x|\omega_2)$.

Step 0: Describe given attributes of problem

- Given: $\vec{x} = [x_1 \dots x_d]$
  - $\Omega = \{\omega_1, \omega_2\}$ (Two classes)
  - $P(\omega_1) = P(\omega_2) = \frac{1}{2}$
  - $p_{ij} = p(x_i = 1 | \omega_j)$
  - $j \in [1,2]$
  - $i \in [d]$
  - $x_i \in \{0,1\}$ Binary (Bernoulli)

- Prove: choose ω_1 if $\sum_{i=1}^{d} x_i > \frac{d}{2}$

$\therefore$ Bayes Decision Rule: choose class with highest posterior probability which leads to the min error probability

Step 1: Start with Bayes Decision Rule to simplify down to a conditional

- Select the maximum posterior to find the minimum probability of error

$\max\{P(\omega_1)P(\vec{x}|\omega_1), P(\omega_2)P(\vec{x}|\omega_2)\}$

- We want to prove ω_1 is max posterior if $\sum_{i=1}^{d} x_i > \frac{d}{2}$

$P(\omega_1)P(\vec{x}|\omega_1) > P(\omega_2)P(\vec{x}|\omega_2)$

- We're also told that $P(\omega_1) = P(\omega_2) = \frac{1}{2}$

$$\frac{\cancel{P(\omega_1)}P(\vec{x}|\omega_1) > \cancel{P(\omega_2)}P(\vec{x}|\omega_2)}{\cancel{P(\omega_1)}}$$

$P(\vec{x}|\omega_1) > P(\vec{x}|\omega_2)$ ← Condition we want

Step 2: Utilizing independence rule write expression out as product (naive bayes)
↳ joint model

$$\prod_{i=1}^{d} p(x_i|\omega_1) > \prod_{i=1}^{d} p(x_i|\omega_2)$$

Step 3: From the joint model, rewrite each distribution over the class variable ω

(i) $\left(\prod_{i=1}^{d} p(x_i|\omega_1)\right)\frac{1}{Z}P(\omega_1) = p(\omega_1|\vec{x})$

$Z = p(x) = \sum_{i=1}^{d} p(\omega_i)p(\vec{x}|\omega_i)$

$= \sum_{i=1}^{d} p(x_i|\omega_i)$ ← NOTE: Same simplification applied to ω_2

(ii) $\sum_{i=1}^{d} p(x_i|\omega_1) > \sum_{i=1}^{d} p(x_i|\omega_2)$

Step 4: Utilizing the last hint to prove "ω_1 if $\sum_{i=1}^{d} x_i > \frac{d}{2}$"

- Given $p_{i1} = p > \frac{1}{2}$

$p_{i2} = 1 - p$

→ Find the posterior for ω_1 & ω_2 (replica for ω_2)

$$P_1(x) = \sum_{i=1}^{d} P(x_i|\omega_1)\,P(\omega_1)$$

$$\frac{P(x|\omega_1)P(\omega_1)}{P(\omega_1|x)} = \sum_{i=1}^{d} P(x_i|\omega_1)P(\omega_1)$$

(i) $$P(\omega_1|x) = \frac{P(x|\omega_1)P(\omega_1)}{\sum_{i=1}^{d} P(x_i|\omega_1)P(\omega_1)} > \frac{1}{2}$$

(ii) $$P(\omega_2|x) = \frac{P(x|\omega_2)P(\omega_2)}{\sum_{i=1}^{d} P(x_i|\omega_2)P(\omega_2)} = 1-p$$

∴ If $P(\omega_1|x) > \frac{1}{2}$ then $\sum_{i=1}^{d} x_i > \frac{d}{2}$ because if the running sum of 1's in the Bernoulli distribution is less than or equal to $d/2$ than $P(\omega_1|x_i=1)$ <u>cannot</u> be the maximum posterior which leads to the minimum probability of error. If it is less than or equal to $d/2$ than there are more 0s in the Bernoulli set which results in ω_2 being the posterior max.

(3) The Ditzler Household Growing up

(a) My parents have two kids now grown into adults. Obviously there is one Greg. I was born on a Weds. What is the probability that I have a brother? You can assume $\mathbb{P}(boy) = \mathbb{P}(girl) = \frac{1}{2}$

Step 1: Calculate prior probabilities

- Given: $\mathbb{P}(boy) = \mathbb{P}(girl) = \frac{1}{2}$
- Priors: boy & boy $- \; P(bb) = (\frac{1}{2})(\frac{1}{2}) = \frac{1}{4}$

  boy & girl $- \; P(bg) = (\frac{1}{2})(\frac{1}{2}) = \frac{1}{4}$

  girl & boy $- \; P(gb) = (\frac{1}{2})(\frac{1}{2}) = \frac{1}{4}$ $\Big\} \; bg + gb = \frac{1}{2}$

  girl & girl $- \; P(gg) = (\frac{1}{2})(\frac{1}{2}) = \frac{1}{4}$ ⟸ Not possible

Step 2: Calculate probability there are two boys given one boy already in the family:

$$P(bb|b) = \frac{P(b|bb)P(bb)}{P(b)} = \frac{(1)(0.25)}{(0.75)} = \frac{1}{3}$$

- $P(b|bb)$ = Probability that both are boys given one is already a boy (1)
- $P(b)$ = Probability of a boy $(\frac{1}{2} + \frac{1}{4})$

$\boxed{P(bb|b) = \frac{1}{3} \text{ or } 0.33 \text{ or } 33\%}$

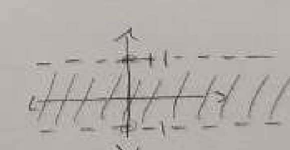
(4) Linear Classifier with a Margin

a) Show that regardless of the dimensionality of the feature vectors, a data set that has just two data points, one from each class, is sufficient to determine the location of the maximum hyperplane. HINT: #1 Consider a data set of two data points, $x_1 \in C_1$ ($y_1 = +1$) and $x_2 \in C_2$ ($y_2 = -1$) and set up the minimization problem (for computing the hyperplane) with appropriate constraints on $w^T x_1 + b$ and $w^T x_2 + b$ and solve it. HINT #2: This can be formed as a constrained optimization problem $\arg\min_{w \in \mathbb{R}^p} \|w\|_2^2$ (subject to some constraint). What is w & b? What are the constraints? How did we solve the constrained optimization problem in Fisher's linear discriminate?

Step 0: Lay out given properties

- Data set $\rightarrow x_1 \in C_1 \; (y_1 = +1)$, $x_2 \in C_2 \; (y_2 = -1)$
- Constraints $\rightarrow w^T x_1 + b = +1$, $\rightarrow w^T x_2 + b = -1$
- optimization $\rightarrow \min \|w\|_2^2$ (minimize square norm)

∴ We want to maximize the margin between the two data points



Step 1: Apply Lagrange multipliers to combined constraints

$y_1(w^T x_1 + b) = 1$

$y_2(w^T x_2 + b) = 1 \rightarrow$ (same as $-y_2(w^T x_2 + b) = -1$)

- Rewriting constraints as function of $g_1(x_1, y_1)$ & $g_2(x_1, y_2)$

$g_1(x_1, y_1) = y_1(w^T x_1 + b) - 1$

$g_2(x_2, y_2) = y_2(w^T x_2 + b) + 1$

- write out Lagrange multipliers

$\mathcal{L}_1(x_1, y_1, \lambda) = f(x_1, y_1) + \lambda \cdot g_1(x_1, y_1)$

$\mathcal{L}_2(x_2, y_2, \lambda) = f(x_2, y_2) + \lambda \cdot g_2(x_2, y_2)$

$\mathcal{L}_1(x_1, y_1, \lambda) = \underset{x \in \mathbb{R}}{\arg\min} \|w\|_2^2 + (y_1(w^T x_1 + b) - 1) \cdot \lambda$

$\mathcal{L}_2(x_2, y_2, \lambda) = \underset{x \in \mathbb{R}}{\arg\min} \|w\|_2^2 + (y_2(w^T x_2 + b) + 1) \cdot \lambda$

— Take the derivative of each Lagrangian function wrt λ and set it to 0

$$\left.\begin{array}{l} \text{(i)}\ \dfrac{d\mathcal{L}_1}{d\lambda} = 0 + y_1 w^T x_1 + y_1 b - 1 = 0 \Rightarrow \underline{y_1(w^T x_1 + b) - 1 = 0} \\ \text{(ii)}\ \dfrac{d\mathcal{L}_2}{d\lambda} = 0 + y_2 w^T x_2 + y_2 b + 1 = 0 \Rightarrow \underline{y_2(w^T x_2 + b) + 1 = 0} \end{array}\right\} \text{original constraints}$$

— Take derivative of each Lagrangian function wrt x & wrt y and set it to 0

(iii) $\dfrac{d\mathcal{L}_1}{dx_1} = 0 + y_1 w^T \lambda = 0 \Rightarrow \underline{y_1 w^T \lambda = 0}$; $\underline{y_1 = 0}$

(iv) $\dfrac{d\mathcal{L}_2}{dx_2} = 0 + y_2 w^T \lambda = 0 \Rightarrow \underline{y_2 w^T \lambda = 0}$; $\underline{y_2 = 0}$

(v) $\dfrac{d\mathcal{L}_1}{dy_1} = 0 + w^T x_1 \lambda + b\lambda \Rightarrow \underline{\lambda w^T x_1 + b = 0}$; $\underline{x_1 = \dfrac{-b\lambda}{w^T\lambda}} = \underline{\dfrac{-b}{w^T}}$

(vi) $\dfrac{d\mathcal{L}_2}{dy_2} = 0 + w^T x_2 \lambda + b\lambda \Rightarrow \underline{\lambda w^T x_2 + b = 0}$; $\underline{x_2 = \dfrac{-b\lambda}{w^T\lambda}} = \underline{\dfrac{-b}{w^T}}$

— Substitute x_1, x_2, y_1, and y_2 back into (i) & (ii)

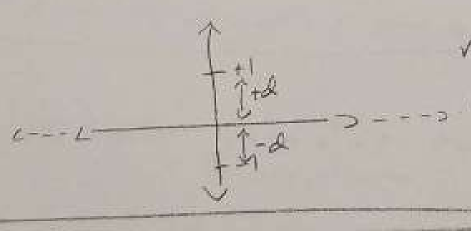
- $0(w^T \cdot (\frac{-b}{w^T}) + b) - 1 = 0$

  $0 - 1 = 0$; ∴ Nothing $= -1$ which is nothing (ie 0)

- $0(w^T \cdot (\frac{-b}{w^T}) + b) + 1 = 0$

  $0 + 1 = 0$; ∴ Nothing $= 1$ which is nothing (ie 0)

∴ Thus the Lagrange multipliers end up being 0 (i.e $\underline{\lambda = 0}$). This means the maximum hyperplane lies on the origin at 0,0 which produces the largest distances from the two data points +1 & −1



maximum hyper plane (ie $y = 0x + 0$)

5) Decision making with Bayes

(A) The Bayes Decision Rule describes the approach we take choosing a class ω for a data point x. This can be achieved modeling $P(\omega|x)$ or $P(x|\omega)P(\omega)/P(x)$. Compare and contrast these two approaches to modeling and discuss the advantages and disadvantages. For the latter model, why might knowing $P(x)$ be useful.

Part 1: $P(\omega|x)$

- $P(\omega|x)$ describes the posterior probability that some event will occur given a second event also occurs. Since we're talking about Bayes Decision Rule, $P(\omega|x)$ represents the maximum posterior probability (aka $\max P(\omega|x)$).
- Advantages of this model is that it describes the bayes decision rule in a more simplistic way (e.g. the probability ω occurs given x also occurs)
- Disadvantages: This model does not describe in details the liklihood, prior, or evidence probabilities (eg $P(x|\omega)$, $P(\omega)$, $P(x)$)
- Compare/contrast: $P(\omega|x)$ can be expanded out to describe $P(x|\omega)P(\omega)/P(x)$, so overall they represent the same results. The contrasting point is how they're expressed (eg $P(x|\omega)P(\omega)/P(x)$ shows liklihood × prior / evidence, whereas $P(\omega|x)$ just shows the posterior probability

Part 2: $P(x|\omega)P(\omega)/P(x)$

- $P(x|\omega)P(\omega)/P(x)$ describes the combination of the sum & product rules to describe bayes. It explicitly shows the liklihood, prior, and evidence probabilities
- Advantage: $P(x|\omega)P(\omega)/P(x)$ is a more verbose way to express the maximum posterior probability (eg $P(\omega|x)$)
- Disadvantage: This model is a little less simplistic to the posterior probability model ($P(\omega|x)$)
- Compare/contrast: Part 1's explanation

Part 3: $P(\omega|x)$ & $P(x|\omega)P(\omega)/P(x)$ w.r.t. Generative, Discriminative, and Discriminate functions

- Generative models: use data available to estimate probabilities of each quantity: priors $p(\omega)$, liklihood $p(x|\omega)$, and evidence $p(x)$. This model is best for the model $P(x|\omega)P(\omega)/P(x)$ => joint distribution $p(x,\omega)$ and normalization to obtain posterior probabilities.
- Discriminitive models: Are based on the posterior probability $p(\omega|x)$ which do not attempt to fully model joint distributions $p(\omega,x)$, but attempts to calculate/estimate posterior directly, thus it is better for the $p(\omega|x)$ models.
  - -> Cannot do outlier detection like generative model
- Discriminate functions: Neither $p(\omega|x)$ or $p(x|\omega)P(\omega)/P(x)$ models this particularly well because these functions focus on decision boundaries which bypass liklihood probabilities & approaches

∴ For the $P(x|\omega)P(\omega)/P(x)$ model describing $P(x)$ can be beneficial because this tells us the probability of event x happening across all other cases (provides us with more beneficial data). $P(x)$ is said to describe the evidence.