भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

# Probabilistic Modelling

Assignment -3 - CS3390

Kartik Srinivas - ES20BTECH11015
October 2, 2022

# Index

# I. Section 3.2

## 1. 3.2.1

We need only show that the baye's optimal we found using generative linear regression can be broken down into a set of $d$ baye's optimal functions, each for one of the labels, but first we find the inverse using Schur's complement lemmas.

$$\hat{f(x)} = \mu_{\mathbf{y}} - \mathrm{A}_{yy}^{-1}\mathrm{A}_{yx}(x - \mu_x) \tag{1}$$

We transform this and write it as

$$\hat{f(x)} = \mu_{\mathbf{y}} + \Sigma_{yx}\Sigma_{xx}^{-1}(x - \mu_x) \tag{2}$$

using the schur complement lemma for inverses, now t eonly thing left to notices is that as the dimension of y actually changes , $\Sigma_{xx}$ will not change, only the term $\Sigma_{yx}$ will get a new row , basically break down the covariance matrix as

$$\begin{bmatrix} f(x)_1 \\ f(x)_2 \\ . \\ . \\ . \\ f(x)_d \end{bmatrix} = \begin{bmatrix} \mu_{y1} \\ \mu_{y2} \\ . \\ . \\ . \\ \mu_{yd} \end{bmatrix} + \begin{bmatrix} \Sigma_{yx_1}K \\ \Sigma_{yx_2}K \\ . \\ . \\ . \\ \Sigma_{yx_d}K \end{bmatrix} \tag{3}$$

where K is the constant vector given by $\Sigma_{xx}^{-1}(x - \mu_x)$

## 2. 3.2.2

In this we only need to multiply them and simplify by completing the square to make this look into terms that are similar to y, in order to show that the marginal is a gaussian, we need to first calculate the joint.

$$\log p(x, y) = -\frac{1}{2}\left[(x - W^T y - b)^T\Sigma_1^{-1}(x - W^T y - b) - (y - \mu_2)^T\Sigma_2^{-1}(y - \mu_2)\right] \tag{4}$$

on simplifying and usin gthe fact that the covariance matirces are symmetric we get

$$\log p(x, y) = -\frac{1}{2}\left[\mathbf{y}^T(\Sigma_2^{-1} + W\Sigma_1^{-1}W^T)\mathbf{y} - 2(x^T\Sigma_1^{-1}W^T + \mu_2\Sigma_2^T - b^T\Sigma_1^{-1}W^T)\mathbf{y} + \ldots\right] \tag{5}$$

The idea is to now complete the squares in the numerator, this hints that the posterior will also be a gaussian

$$\frac{1}{2}z^T Az + b^T z + c = \frac{1}{2}(z + A^{-1}b)^T(A)(z + A^{-1}b) + c - \frac{1}{2}b^T A^{-1}b \quad (6)$$

We get that the mean of the posterior will correspond to the $A^{-1}b$ term and the precision matrix will be 'A'

$$\boxed{A = \Sigma_{y|x}^{-1} = (\Sigma_2^{-1} + W\Sigma_1^{-1}W^T)}$$

$$A^{-1}b = \mu_{y|x} = (\Sigma_{y|x})((x^T - b^T)\Sigma_1^{-1}W^T + \mu_2^T\Sigma_2^{-1})^T)$$

$$\therefore \boxed{\mu_{y|x} = (\Sigma_{y|x})(W\Sigma_1^{-1}(x - b) + \Sigma_2^{-1}\mu_2)}$$

## 3. 3.2.3

# II. Section 3.3

## 1. Murphy - 3.20

### i. Part(a)

If we take the distribution of the binary vectors separately, then we are forced to assign a probability distribution over all the $2^D$ vectors for a particular $y = c$ Since there are only finitely many , we can use a categorical for every class over all the vectors

$$p(x|y = c) = \mathbf{Cat}(x|\theta_\mathbf{c}) \quad \theta_c \in \mathbb{R}^{2^D} \quad \sum(\theta_\mathbf{c}) = 1 \tag{7}$$

Number of parameters $= (2^D - 1) \times C$

### ii. Part(b)

If the sample size is large, the case with large number of parameters (non Naive- modelling) will be able to perform better than the Naive classifier, because the Non-Naive Model is a superset of the Naive model, and therefore the Non-Naive model will perform better on training of parameters through a sufficient number of examples.

### iii. Part (c)

If the number of examples are insufficient, it will be very difficult to train the large number of parameters of the Non-Naive case. However the Naive case has lesser parameters and hence the model will be able to learn better with lesser number of samples to work with.

### iv. Part (d), (e)

## 2. Murphy - 3.22

We only need to take the fraction of the examples where that particular word in the vocabular occurs (since this is a Bernoulli Model, these will be

the likelihood estimates)

$$\text{Total examples} = 3 + 4 = 7 \tag{8}$$
$$\theta_{non-spam} = 4/7 \tag{9}$$
$$\theta_{spam} = 3/7 \tag{10}$$
$$\theta_{secret|spam} = 2/3 \tag{11}$$
$$\theta_{dollar|spam} = 1/3 \tag{12}$$
$$\theta_{sports|non-spam} = 2/4 \tag{13}$$
$$\theta_{secret|non-spam} = 1/4 \tag{14}$$

## 3. Murphy 4.18

$$\theta = 0.5, 0.5, 0.5 \qquad \mu = \mu_c, \sigma = \sigma_c \quad \forall c$$

See that all the classes model the same Normal and Bernoulli Distributions!, the naive baye's classifier ( which assumes conditional independence) will not be actually say anything more about the conditional $p(y|x)$ than what p(y) would say(the prior is the only way to judge an example, Since a single value of x are all equally weighted by the class conditionals.)

$$p(y_i|x) = \frac{p(x_1|y_i)p(y_i)}{\sum p(x_1|y_j)p(y_j)} = \frac{p(y_i)}{\sum(p(y_i))} = p(y_i) \tag{15}$$

So for all of them the answer will be

$$p(y|x_1 = 0) = p(y|x_2 = 0) = p(y|x_1, x_2 = 0) = \pi = [0.5, 0.25, 0.25] \tag{16}$$

## 4. Murphy 4.19

The expression will not boil down much, the determinants of the covariances will be related though, this will simplify the expression (Uniform prior has been assumed)

$$p(y=1|x) = \frac{p(y=1,x)}{\sum p(x=i|y)p(y)}$$

$$p(y=1|x) = \frac{1/|\Sigma_1|^{1/2}}{1/|\Sigma_1|^{1/2}\mathbf{Gau}(x|y=1) + 1/|\Sigma_0|^{1/2}\mathbf{Gau}(x|y=0)}$$

$$|\Sigma_1| = k^d|\Sigma_0| \quad \& \quad \Sigma_1^{-1} = \frac{\Sigma_0^{-1}}{k}$$

$$f_c(x) = -\frac{1}{2}(x-\mu_c)^T\Sigma_c(x-\mu_c)$$

$$p(y=1|x) = \frac{1}{1+k^{d/2}\exp(f_0(x)-f_1(x))}$$

$$p(y=1|x) = \frac{1}{1+\exp(f_0(x)-f_1(x)+d/2\ln(k))}$$

$$p(y=1|x) = \frac{1}{1+\exp(f_0(x)-f_1(x)+d/2\ln(k))}$$

$$p(y=1|x) = \frac{1}{1+\exp\left(-\frac{1}{2}(x^TAx+bx+C)\right)}$$

$$A = \frac{k-1}{2k}(\Sigma_0^{-1})$$

$$b = \frac{2}{k}(k\mu_0^T - \mu_1^T)(\Sigma_0)^{-1}$$

Geometric interpretation If the value of k increases the probability of lying in the 1st class decreases , (more spread out, so inorder to compensate for this the decision surface will now move closer to the mean of the class - 1 distribution.

Decision surface:- I believe the decision surface will be some sort of hyperbolic surface, the Mahalanobis distances when subtracted yield a constant.

$$(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_1)^T - \frac{1}{k}(x-\mu_1)^T\Sigma_0^{-1}(x-\mu_1)^T = \frac{d}{2}ln(k) \tag{17}$$

## 5. Problem 4.21

The decision region is just the equality of the probabilities of the class conditionals

# III. Shalev 24.2

The situation described is MLE being performed on m samples that have supposedly been picked from a bernoulli distribution basically the regularized version adds one positive example($x_i = 1$) and one negtive example ($x_i = 0$) this is done to prevent overfitting in the case that all the results in the samples are ($x_i = 1$), where normal MLE would have yielded $\theta = 1$. Therefre in-order to avoid this scenario,(where $\theta_{MLE} = 0, 1$) we add two extra terms supporting each class. we already know that the solution for MLE is the fraction of positive samples within the set of samples, with one extra positive sample added.

$$\text{Num of positive samples} = \sum_{i=1}^{i=m} x_i \ + 1$$

$$\text{Total samples} = m + 1 + 1$$

$$\theta_{MLE} = \text{Fraction of 'heads'} = \ \frac{\sum x_i + 1}{m + 2}$$

# IV. Murphy 3.6

It is better to deal with the log likelihood in this case, once we multiply all the likelihoods and take the logarithm, we obtain the following expression

$$\frac{\partial}{\partial \lambda} \log\Big(\prod \text{Poi}(x_i|\lambda_i)\Big) = 0 \tag{18}$$

$$\frac{\partial}{\partial \lambda}(-\lambda m + \sum x_i \log \lambda - \sum \log x_i!) = 0 \tag{19}$$

$$\therefore \lambda = \frac{\sum x_i}{m} \tag{20}$$

This also makes sense , the poisson is an approximated version of the binomial, the average number of successes, is just the average of the success rate.

# V. Murphy 3.8

The uniform distribution will try to keep all the points 'just' inside the rectangular bump, i.e. we must minimize a such that all the points $x_i \in$

$[-a, a]$.(because the probability of a sample being inside is non zero and $\propto 1/a$)

$$\hat{a} = max(|x_i|) \tag{21}$$

$$p(x_{n+1}|\hat{a}) = \frac{1}{2\hat{a}} I(|x_{n+1}| \leq \hat{a})) \tag{22}$$

There are 2 problems with this approach, one is that it does not consider anything that is even slightly outside the maximum absolute value seen in the training data. The model by selection itself has become prone to a large generalization error. Secondly, think about the scenario where we have a single misclassified sample. This misclassified sample with a large value of x will offset the entire training procedure , and $\hat{a} = |x_{misclassified}|$. The model has no flexibility, it only cares about a single training sample

# VI. Murphy 3.11

Multiplying all the probabilities and taking the log likelihood we get

$$\frac{\partial}{\partial\theta}[mlog(\theta) - \theta(\sum x_i)] = 0 \tag{23}$$

$$\hat{\theta} = \frac{m}{\sum x_i} \tag{24}$$

For the example given

$$\hat{\theta} = \frac{3}{5+6+4} = 0.2 \tag{25}$$

# VII. Murphy - 4.5

The posterior has been given to us, and we are to construct the joint.

| | $y=0$ | $y=1$ |
|---|---|---|
| $x=0$ | $\theta_2(1-\theta_1)$ | $(1-\theta_2)(1-\theta_1)$ |
| $x=1$ | $(1-\theta_2)\theta_1$ | $(\theta_1)(\theta_2)$ |

Samples are i.i.d form the joint. What is to notice is that the distributions $p(y|x=0)$ and $p(y|x=1)$ are tied using Bernoulli's that have complementary parameters. On Multiplication of the joint probabilities from the dataset we

get

$$p(D|\theta) = \theta_1^4\theta_2^4(1-\theta_1)^3(1-\theta_2)^3 \tag{26}$$

$$\text{Maximizing we get } \theta_1 = \theta_2 = \frac{4}{7} \tag{27}$$

$$p(D|\theta) = \frac{4^8 3^6}{7^{14}} \tag{28}$$

# VIII. Murphy 10.1

We must write this neatly, **w** is a matrix and $x_i$ is a vector. $\mu_{ik} = \text{Softmax}(w^T x_i)_k$ SInce we are only dealing wit the i'th elements here we can drop the subscript and note that $\eta_{ij} = (w^T x_i)_j$

$$\frac{\partial \mu_{ik}}{\partial \eta_{ij}} = \frac{\partial}{\partial (w^T x_i)_j}(\frac{e^(w^T x_i)_k}{\sum_l (w^T x_i)_l}) \tag{29}$$

$$\tag{30}$$

$$= \frac{I(k=j)e^{(w^T x_i)_k}\sum_l e^{(w^T x_i)_l} - e^{(w^T x_i)_j}e^{(w^T x_i)_k}}{(\sum_l e^{(w^T x_i)_l})^2} \tag{31}$$

$$= \mu_{ik}(I(k=j) - \mu_{ij}) \tag{32}$$

Calculating the Hessian.

# IX. Murphy 10.3

LDA models linear parameter scores using MLE on the joint distribution, QDA is more general than LDA , ie QDA contains LDA as a apecial case, so the error that is given (in terms of posterior likelihood over the training set) for QDA will be always better than that of LDA. Similar arguments can be given for QuadLog and LinLog. QuadLog will contain LinLog and hence the log likelihood on the dataset for QuadLog will atleast be as good as that of LinLog.

$$L(\text{QuadLog}) \leq L(\text{LinLog}) \tag{33}$$

$$L(\text{QuadLog}) \leq L(\text{LDA(GaussI)}) \tag{34}$$

$$L(\text{QDA(GuassX)}) \leq ? \geq L(\text{QuadLog}) \tag{35}$$

$$\tag{36}$$

# X. Section 3.4

## 1. Murphy - 7.5

This can be viewed in 2 ways , both function or discriminative modelling. Both methods yield the same answer

$$\frac{\partial}{\partial w_0}(\sum_{i=1}^{i=m}(y_i - w^T x_i - w_0)^2) = 0 \tag{37}$$

$$\therefore \hat{w_0} = \frac{1}{m}y_i - w^T\frac{1}{m}(\sum \mathbf{x_i}) = \bar{y} - w^T\bar{x} \tag{38}$$

$$\frac{\partial}{\partial \mathbf{w}}(\sum_{i=1}^{i=m}(y_i - w^T x_i - \hat{w_0})^2) = 0 \tag{39}$$

$$\frac{\partial}{\partial w}(\sum_{i=1}^{i=m}(y_i - \bar{y} - w^T(x_i - \bar{x}))^2) = 0 \tag{40}$$

The problem is not just standard linear regression( with the shifted y's and x's)!. We know that solving (40) will yield the equation $\bar{X}\bar{X}^T w = \bar{X}\bar{y}$ (Where the examples are along the **columns** not rows). Provided that the Row rank of X is full , the solution to (40) will be $\mathbf{w} = (\bar{X}\bar{X}^T)^{-1}\bar{X}\bar{y}$

$$\bar{X}, \bar{y} = \textbf{columns arranged} \quad x_i - \bar{x}, y_i - \bar{y}$$

## 2. Murphy 7.6

We can obtain this equation by directly solving [40] and treating $\mathbf{w} = w$ on differentiating we obtain

$$\sum(y_i - \bar{y} - w(x_i - \bar{x_i}))(x_i - \bar{x_i}) = 0 \tag{41}$$

$$\therefore w = \frac{\sum(y_i - \bar{y})(x - \bar{x})}{\sum(x_i - \bar{x})^2} \tag{42}$$

This makes complete sense , the numerator is a measure of the linear relationship between x and y(rate of comparative growth) $\mathbf{cov}(X, Y)$ and the denominator is only a normalization term of x, the variance. So one we multiply $wx$ we get the value of y, only shifted by a certain value in -order to compensate for this, we get the bias term $w_0 = \bar{y} - \bar{w}\bar{x}$

## 3. Murphy 7.9

We are basically asked to find out the values of the mean and the posterior in generative modelling and to find out the relationship between both the

modelling strategies. We know from generative modelling and applying schur complement lemma we obtain the following

$$\mu_{y|x} = f(x) = \mu_y + \hat{\Sigma_{yx}}\hat{\Sigma_{xx}}^{-1}(x - \mu_x)$$

Let us simplify this using our notation earlier,

$$\mu_{y|x} = \mu_y + \left[\frac{1}{m}\sum(y - \bar{y})(x - \bar{x})^T\right](\bar{X}\bar{X}^T)^{-1}(x - \mu_x) \tag{43}$$

$$\textbf{On Simplifying we get} \quad \mu_{y|x} = \mu_y + ((\bar{X}\bar{X}^T)^{-1}\bar{X}\bar{y})^T(x - \bar{x}) \tag{44}$$

$$\textbf{Using the solution to (40)} = \bar{y} - w^T\bar{x} + w^T x \tag{45}$$

$$= w_0 + w^T\bar{x} \tag{46}$$

This shows that both the models are equivalent in terms of modelling capability, however the advantages of using the second are that we now have the joint distribution $p(y, x)$ sampling from this will allow us to generate prospective samples, this is not there with the posterior. The disadvantage is that we end up applying Law of Large Numbers for too many variables when doing generative modelling, but the same cannot be said for discriminative. Discriminative modelling will give us better PAC bounds than generative ones.

# XI.