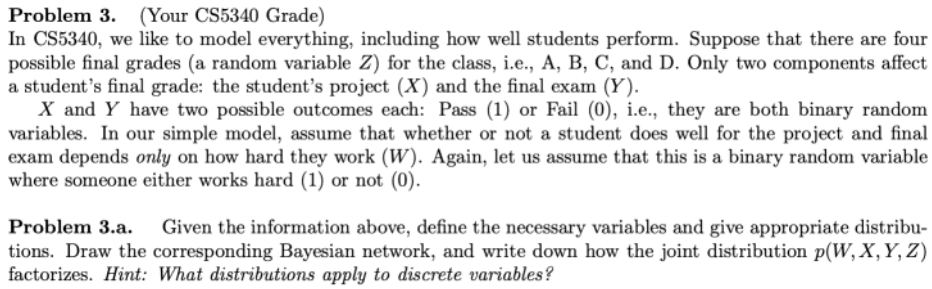
Problem 3

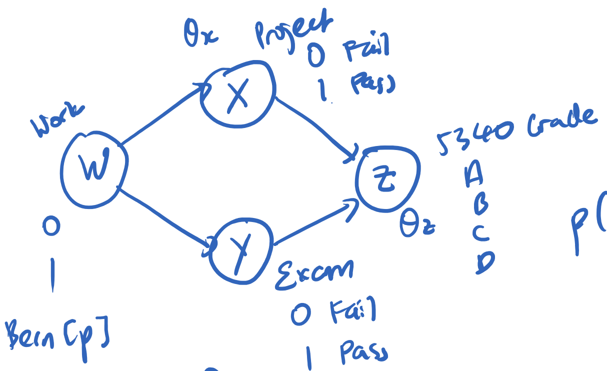
Thursday, 2 February 2023 8:33 AM



Problem 3. (Your CS5340 Grade)
In CS5340, we like to model everything, including how well students perform. Suppose that there are four possible final grades (a random variable Z) for the class, i.e., A, B, C, and D. Only two components affect a student's final grade: the student's project (X) and the final exam (Y).

X and Y have two possible outcomes each: Pass (1) or Fail (0), i.e., they are both binary random variables. In our simple model, assume that whether or not a student does well for the project and final exam depends *only* on how hard they work (W). Again, let us assume that this is a binary random variable where someone either works hard (1) or not (0).

Problem 3.a. Given the information above, define the necessary variables and give appropriate distributions. Draw the corresponding Bayesian network, and write down how the joint distribution $p(W, X, Y, Z)$ factorizes. *Hint: What distributions apply to discrete variables?*



X	W	$p(x \mid w)$ $\neq p(x,w)$
0	0	θ_{00} ✓
0	1	θ_{01}
1	0	θ_{10} ✓
1	1	θ_{11}

$p(X=0 \mid W=0)$

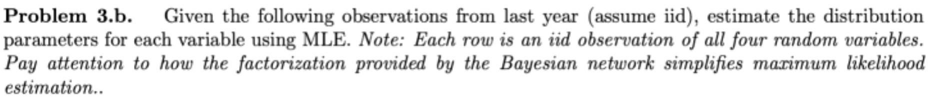
$p(X=1 \mid W=0)$

$$p(W,X,Y,Z)$$
$$= p(W)p(X \mid W)p(Y \mid W)p(Z \mid X,Y)$$

θ_w θ_y

$p(\underline{\underline{w}})$ $p(Y \mid W)$

$p(w\,;\theta_w)$



Problem 3.b. Given the following observations from last year (assume iid), estimate the distribution parameters for each variable using MLE. *Note: Each row is an iid observation of all four random variables. Pay attention to how the factorization provided by the Bayesian network simplifies maximum likelihood estimation..*

$p(x \mid \theta)$ ← MAP Bayes

$p(x\,;\theta)$ MLE

MLE.

$$\arg\max_{\theta} \lg \underbrace{p(D \mid \theta_x, \theta_y, \theta_w, \theta_z)}$$

$$\lg \prod_i p(w_i, x_i, y_i, z_i\,; \theta_x, \theta_y, \theta_w, \theta_z)$$

$\theta = \{\theta_x, \theta_y, \theta_w, \theta_z\}$

$$\sum_i \lg \underbrace{p(w_i, x_i, y_i, z_i\,; \theta_x, \theta_y, \theta_w, \theta_z)}$$

$$\lg p(w_i\,;\theta) + \lg p(x_i \mid w_i\,;\theta) + \lg p(y_i \mid w_i\,;\theta) + \lg p(z_i \mid x_i, y_i\,;\theta).$$

$$\arg\max_{\theta_w, \theta_x, \theta_y, \theta_z} \sum_i \lg p(w_i\,;\underline{\underline{\theta_w}}) + \underline{\lg p(\underline{x_i} \mid \underline{w_i}, \theta_x) + \lg p(y_i \mid w_i\,;\theta_y) + \lg p(z_i \mid x_i, y_i, \theta_z)}$$

Find θ_w. only need to consider $\sum_i \lg p(w_i\,;\theta_w)$.

Find θ_x.

X	W	$p(x \mid w)$

Find θ_x.

X	W	p(x|w)	
0	0	θ_{00}	θ_x
0	1	θ_{01}	
1	0	θ_{10}	
1	1	θ_{11}	

$$\lg p(x|w=0) = \sum_i \prod_j \theta_{j0}^{[x_i=j]}$$

$$\theta^x (1-\theta)^{1-x}$$

argmin $\mathcal{L}$ where

$$\mathcal{L} = -\sum_i \sum_j [x_i=j] \lg \theta_{j0} + v\left[\sum_j \theta_{j0} - 1\right]$$

$$\frac{\partial \mathcal{L}}{\partial \theta_{j0}} = 0 \qquad \frac{\partial \mathcal{L}}{\partial v} = 0 \leftarrow$$

$$\Rightarrow \boxed{\theta_{j0} = \frac{N_j}{\sum_k N_k}}$$

N_j: # times we observe $x_i = j$

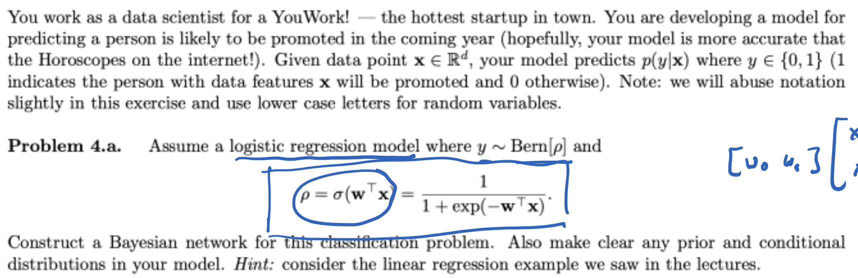
$\sum_k N_k$: normalizer.

Problem 4

Thursday, 2 February 2023 8:36 AM

https://www.chinahighlights.com/travelguide/chinese-zodiac/monthly-fortune-for-rooster.htm

Note: I do not believe in horoscopes nor am suggesting that you should!

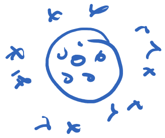


You work as a data scientist for a YouWork! — the hottest startup in town. You are developing a model for predicting a person is likely to be promoted in the coming year (hopefully, your model is more accurate that the Horoscopes on the internet!). Given data point $\mathbf{x} \in \mathbb{R}^d$, your model predicts $p(y|\mathbf{x})$ where $y \in \{0,1\}$ (1 indicates the person with data features $\mathbf{x}$ will be promoted and 0 otherwise). Note: we will abuse notation slightly in this exercise and use lower case letters for random variables.

Problem 4.a. Assume a logistic regression model where $y \sim \text{Bern}[\rho]$ and

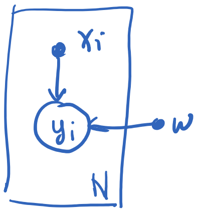
$$\boxed{\rho = \sigma(\mathbf{w}^\top\mathbf{x}) = \frac{1}{1+\exp(-\mathbf{w}^\top\mathbf{x})}.}$$

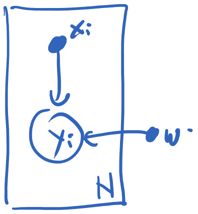
Construct a Bayesian network for this classification problem. Also make clear any prior and conditional distributions in your model. *Hint:* consider the linear regression example we saw in the lectures.



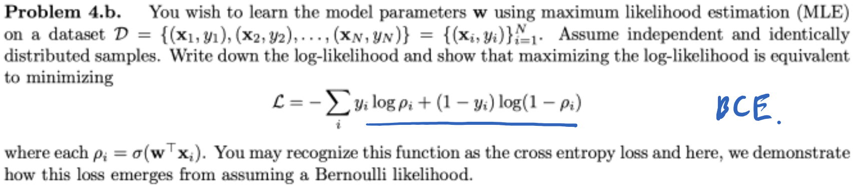
$$[w_0 \; w_1]\begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = w_0 x_0 + w_1 x_1$$

$$x \rightarrow \varphi(x).$$





$$p(y_i|x_i, w) = \text{Bern}[\rho_i]$$
$$= \rho_i^{y_i}(1-\rho_i)^{1-y_i} \;\|$$
$$\rho_i = \sigma(w^T x)$$



Problem 4.b. You wish to learn the model parameters $\mathbf{w}$ using maximum likelihood estimation (MLE) on a dataset $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)\} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$. Assume independent and identically distributed samples. Write down the log-likelihood and show that maximizing the log-likelihood is equivalent to minimizing

$$\mathcal{L} = -\sum_i \underline{y_i \log \rho_i + (1-y_i)\log(1-\rho_i)} \qquad \text{BCE.}$$

where each $\rho_i = \sigma(\mathbf{w}^\top\mathbf{x}_i)$. You may recognize this function as the cross entropy loss and here, we demonstrate how this loss emerges from assuming a Bernoulli likelihood.

$$\underset{w}{\text{argmax}} \log \prod_i^N p(y_i|x_i, w)$$

$$= \underset{w}{\text{arg min}} \underbrace{-\log \prod_i^N p(y_i|x_i, w)}_{\mathcal{L}}$$

$$\mathcal{L} = -\sum_i^N \lg p(y_i|x_i, w)$$

$$= -\sum \lg \rho_i^{y_i}(1-\rho_i)^{1-y_i}$$

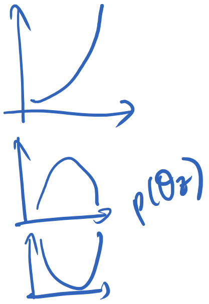
$$= -\sum y_i \lg \rho_i + (1-y_i)\lg(1-\rho_i) \quad //$$

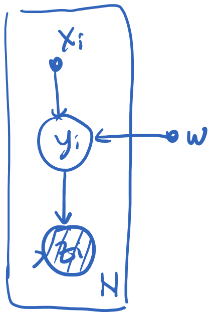
Problem 4.c. During inspection of your training data, you find that some of the data points are mislabelled! You could look through the data manually to find the mislabelled data but this seems rather labor intensive. Can you adjust your model to account for the wrong labels? Let us introduce a new random variable z which represents the observed (possibly wrong) label. The actual y is now hidden (or "latent"). You know that the variables are related via the conditional distribution,

z	y	$p(z\|y)$
0	0	0.75
0	1	0.05
1	0	0.25
1	1	0.95

$\leftarrow \theta_z$

Given this information, design a new Bayesian network (*hint:* extend the basic classification model with z) and derive the log-likelihood (*Hint:* recall the sum rule). How is the new MLE optimization function different from the one you derived in the previous subsection?





$$p(y_i, z_i \mid x_i, w)$$

$$= p(z_i \mid y_i)\, p(y_i \mid x_i, w)$$

$$\underset{w}{\arg\max} \log \prod_i^N p(z_i \mid x_i, w)$$

$$\mathcal{L} = -\sum_i \lg \underline{p(z_i \mid w, x_i)}$$

$$= -\sum_i \lg \prod_{j\in\{0,1\}} \theta_j^{[z_i=j]}$$

$$= -\sum_i z_i \lg \theta_1 + (1-z_i)\lg\theta_0$$

z	$p(z_i \mid w, x_i)$	
0	θ_0	
1	θ_1	$(1-\theta_0)$

Bernoulli

$$p(\underline{z_i} \mid w, x_i)$$

$$= \sum_{y_i} p(y_i, \underline{z_i} \mid w, x_i)$$

$$= \underline{\sum_{y_i} p(z_i \mid y_i)\, p(y_i \mid x_i, w)}$$

$$\mathcal{L} = -\sum_i \underline{z_i} \lg(0.95\rho_i + 0.25(1-\rho_i)) + (1-z_i)\lg(0.05\rho_i + 0.75(1-\rho_i))$$

$$\rho_i = \sigma(w^T x_i)$$

$$\theta_1 = p(z_i = 1 \mid w, x_i)$$

$$= \underline{p(z_i=1 \mid y_i=1)}\, \overbrace{p(y_i=1 \mid x_i, w)}^{\rho_i} + p(z_i=1 \mid y_i=0)\, p(y_i=0 \mid x_i, w).$$

$$\theta_0 = p(z_i=0 \mid y_i=1)\, p(y_i=1 \mid x_i, w) +$$

$p(z_i=0 | y_i=0)\, p(y_i=0 | x_i, w)$.

model → likelihood → $\underline{\underline{\text{loss}}}$ / obj function.