COL774

Machine Learning

Solutions (Minor)

Q.2. Poisson Distribution:- $y \sim \text{Poisson}(\lambda)$

(a) $\Rightarrow P(y=k;\lambda) = \dfrac{\lambda^k e^{-\lambda}}{k!}$

$$\Rightarrow p(y) = \frac{\lambda^y e^{-\lambda}}{y!} \quad -(1)$$

Now, if $y \sim \text{exp-family}(\eta)$

$$\Rightarrow P(y;\eta) = b(y)\, e^{(\eta y - a(\eta))} \quad -(2)$$

In (1), we can write

$$\log [P(y;\lambda)] = y \log \lambda - \lambda - \log y! \quad -(3)$$

Taking log in (2) we get

$$\log P(y;\eta) = \log b(y) + \eta y - a(\eta) \quad -(4)$$

Equating (3) & (4) we get

$$\eta = \log \lambda \quad -(a)$$

$$a(\eta) = \lambda = e^{\eta} \quad -(b)$$

$$\log b(y) = -\log y! \Rightarrow b(y) = \frac{1}{y!} \quad -(c)$$

$\Rightarrow$ $y \sim \text{Poisson}(\lambda)$ belongs to exponential family

$y = \theta^T x \Rightarrow \log \lambda = \theta^T x \Rightarrow \lambda = e^{\theta^T x}$

(b) $\log p(y;\eta) = \log b(y) + \eta y - a(\eta)$
substituting values for $b(y)$ & η we get:-

$$\sum_{i=1}^{m} \log p(y^{(i)};\theta) = \sum_{i=1}^{m} \left[-\log(y^{(i)}!) + \theta^T x^{(i)} y^{(i)} - e^{\theta^T x^{(i)}} \right]$$

$$= LL(\theta)$$

$$\Rightarrow \nabla_\theta LL(\theta) = \sum_{i=1}^{m} 0 + x^{(i)} y^{(i)} - e^{\theta^T x^{(i)}} x^{(i)}$$

$$= \sum_{i=1}^{m} \left[\boxed{y^{(i)}} - e^{\theta^T x^{(i)}} \right] x^{(i)}$$

(c) $$\nabla_\theta LL(\theta) = \sum_{i=1}^{m} \left(y^{(i)} - e^{\theta^T x^{(i)}} \right) x^{(i)}$$

$$\Rightarrow \nabla^2_\theta LL(\theta) = \sum_{i=1}^{m} \left[- e^{\theta^T x^{(i)}} \right] x^{(i)} x^{(i)T}$$

$$\left[\because \frac{\partial LL(\theta)}{\partial \theta_j} = \sum_{i=1}^{m} \left(y^{(i)} - e^{\theta^T x^{(i)}} \right) x_j^{(i)} \right.$$

$$\Rightarrow \frac{\partial^2 LL(\theta)}{\partial\theta_j \partial\theta_k} = \sum_{i=1}^{m} \left(-e^{\theta^T x^{(i)}}\right) x_j^{(i)} x_k^{(i)}$$

$$\Rightarrow \text{Hessian matrix}: H = \sum_{i=1}^{m} \left(-e^{\theta^T x^{(i)}}\right) x^{(i)} x^{(i)T}$$

$Z \in \mathbb{R}^n$

$$\Rightarrow Z^T H Z = \sum_{i=1}^{m} Z^T x^{(i)} x^{(i)T} Z \left[-e^{\theta^T x^{(i)}}\right]$$

$$= \sum_{i=1}^{m} \underbrace{\left[Z^T x^{(i)}\right]^2}_{\geq 0} \underbrace{\left[e^{\theta^T x^{(i)}}\right]}_{>0} (-1)$$

$\Rightarrow Z^T H Z \leq 0 \Rightarrow H$ is $-$ve semi-definite

$\Rightarrow LL(\theta)$ is concave function of θ.

Hence proved

Q.2. (i) The model is overfitting. The improvement advice would be to try a simpler model first. Alternatively, if additional data is available they can train original model with additional data to see if generalization accuracy improves.

(ii) The model seems to be doing well in both training as well as generalization ... further

Student can try to improve things further by trying a more sophisticated model for possible improvement in train/val accuracy. This might require training with additional data to prevent overfitting

(III) Both training & validation accuracy are low. Model is underfitting & student should increase the complexity of their model to better represent patterns in underlying data.

(b) $$\theta_{j|y=k} = \frac{\sum_{i=1}^{m} 1\{y^{(i)}=k\} 1\{x_j^{(i)}=1\} + C \cdot 1}{\sum_{i=1}^{m} 1\{y^{(i)}=k\} + C \cdot 2}$$

C controls overfitting because it does not allow the model to overfit the training data in extremely noisy scenarios (by having a prior). For example if frequency count of some word w_e is 0 in training for some class $y=k$, the introduction of C term makes sure that the probability of $(y=k) \neq 0$ even if w_e does not appear in the document. Similarly, if frequency count of w_ℓ is very very low compared to other words, for some class $y=k$, the introduction of c term (smoothing) ensures that the probability is increased by adding a prior in form of c term in numerator (& denominator). Larger the value of c, larger the impact of this term is. Overfitting is prevented by increasing the effective count for each attribute (w_ℓ) by c, for each class $y=k$, by effectively adding a document in which each word we appears c times. By carefully choosing this mechanism, a uniform prior over words (attributes) is introduced reducing overfitting.

Large value of c :- Model is too heavily biased by prior & it will tend to underfit

small value of c :- Prior is very weak & model might start overfitting.

Q.3.

(a)

variances of individual components

$$\Sigma_1 = \begin{bmatrix} \sigma_{11}^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_{1n}^2 \end{bmatrix}$$

Assume (wlog) $x^{(i)} \in \mathbb{R}^n$

cross diagonal entries are zero, since attribute are independent (given the class) & hence un-correlated with each other. (i.e. covariance terms will be zero).

Similarly $\hat{\Sigma}_2 = \begin{bmatrix} \sigma_{21}^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_{2n}^2 \end{bmatrix}$

$\log\left[\prod_{i=1}^{m} p(y^{(i)}, x^{(i)}; \theta)\right]$

$= \sum_{i=1}^{m} \log\left[p(y^{(i)}; \phi) \left[p(x^{(i)} | y^{(i)}; \theta)\right]\right]$

$= \sum_{i=1}^{m} 1\{y^{(i)} = k\}[\phi_k] + \log \frac{1}{(2\pi)^{n/2} |\hat{\Sigma}_{y^{(i)}}|^{1/2}} e^{-\frac{[x^{(i)} - \mu_{y^{(i)}}]^T \hat{\Sigma}_{y^{(i)}}^{-1} (x^{(i)} - \mu_{y^{(i)}})}{2}}$

Since $\hat{\Sigma}_{y^{(i)}}$ is diagonal $\forall (i)$ $\{y^{(i)} \in \{1 \dots r\}\}$

we can decompose above expression as:- $(\mu_k \equiv \mu_{k1} \dots \mu_{kn})$

$= \sum_{i=1}^{m} \sum_{k=1}^{r} 1\{y^{(i)} = k\} \log \phi_k + \sum_{i=1}^{m} \left[\log \prod_{j=1}^{n} \left[\frac{1}{(2\pi)\sigma_{y^{(i)}j}^2}\right] + \log e^{\sum_{j=1}^{n} -\left[\frac{(x_j^{(i)} - \mu_{y^{(i)}j})^2}{2\sigma_{y^{(i)}j}^2}\right]}\right]$

$= \sum_{i=1}^{m} \sum_{k=1}^{r} 1\{y^{(i)} = k\} \log \phi_k + \sum_{i=1}^{m} \left[\sum_{j=1}^{n} \log \frac{1}{\sqrt{2\pi \sigma_{y^{(i)}j}^2}}\right.$

$+ \sum_{j=1}^{n} (-1/2) \left[\left(x_j^{(i)} - \mu_{y^{(i)}j}\right)^2 \dots \right]\Big]$

Part (b) $\sum_{j=1}^{r} (1/2) \left[\frac{\ldots}{\sigma_{y^{(i)}j}^2} \right] \Big]$

(i) Differentiating wrt μ_{kj}

$$\frac{\partial LL(\theta)}{\partial \mu_{kj}} = 0 + 0 + \sum_{i=1}^{m} (-1/2) \sum_{k=1}^{r} 1\{y^{(i)}=k\} \frac{\partial}{\partial \mu_{kj}} \left[\frac{(x^{(i)}_j - \mu_{kj})^2}{(\sigma_{kj})^2} \right]$$

$$= \sum_{i=1}^{m} (-1/2) \sum_{k=1}^{r} 1\{y^{(i)}=k\} \, 2 \frac{(x^{(i)}_j - \mu_{kj})}{(\sigma_{kj})^2}$$

Equating to zero, we get

$$\frac{\partial LL(\theta)}{\partial \mu_{kj}} = \sum_{i=1}^{m} \left[\sum_{k=1}^{r} 1\{y^{(i)}=k\} (x^{(i)}_j - \mu_{kj}) \right] = 0$$

$$\Rightarrow \mu_{kj} = \frac{\sum_{i=1}^{m} 1\{y^{(i)}=k\} x^{(i)}_j}{\sum_{i=1}^{m} 1\{y^{(i)}=k\}}$$

$$\Rightarrow \mu_k = \frac{\sum_{i=1}^{m} 1\{y^{(i)}=k\} x^{(i)}}{\sum_{i=1}^{m} 1\{y^{(i)}=k\}}$$

$\Rightarrow$ which is same as derived in class.

(ii) Differentiating wrt σ_{1d}, we get

$$\frac{\partial LL(\theta)}{\ldots} = \frac{\partial}{\ldots} \left[\sum_{i=1}^{m} 1\{y^{(i)} \in \{1, 3, \ldots, r-1\}\} \sum^{n} \ldots x^{(i)} - \mu_{y^{(i)}} \ldots \right.$$

$$\frac{\partial}{\partial \sigma_{1j}} \ldots \sum_{j=1}^{n} \log \frac{1}{\sqrt{(2\pi\sigma_{1j}^2)}} + \sum_{j=1}^{n} - \frac{(x_j - \mu_{1j})^2}{2\sigma_{1j}^2}$$

$$= \frac{\partial}{\partial \sigma_{1j}} \sum_{i=1}^{m} \Bigg[1\{y^{(i)} \text{ is odd}\} \cdot \sum_{j=1}^{n} \log \frac{1}{\sqrt{2\pi}} - \log(\sigma_{1j}) + \sum_{j=1}^{n} - \frac{(x_j^{(i)} - \mu_{y^{(i)}j})^2}{(2\sigma_{1j}^2)} \Bigg]$$

$$= \sum_{i=1}^{m} 1\{y^{(i)} \text{ is odd}\} \Bigg[\frac{(-)}{(\sigma_{1j})} + \frac{(x_j^{(i)} - \mu_{y^{(i)}j})^2}{\cancel{2}(\sigma_{1j})^3} \cdot \cancel{2} \Bigg]$$

Equating to zero we get:-

$$\sigma_{1j}^2 = \frac{\sum_{i=1}^{m} 1\{y^{(i)} \text{ is odd}\} (x_j^{(i)} - \mu_{y^{(i)}j})^2}{\sum_{i=1}^{m} 1\{y^{(i)} \text{ is odd}\}}$$

$\Rightarrow$ Empirical variance computed across class which share the Σ parameter

$\Rightarrow$ same as expression derived in the

Note:- [class] expression derived in class was:- (equivalent)

for the general case

$$\hat{\Sigma}_1 = \frac{\sum_{i=1}^{m} 1\{y^{(i)} \text{ is odd}\} (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T}{\sum_{i=1}^{m} 1\{y^{(i)} \text{ is odd}\}}$$

Since cross diagonal entries are zero (due to naive Bayes assumption)

we get $$\sigma_{1j}^2 = \frac{\sum_{i=1}^{m} 1\{y^{(i)} \text{ is odd}\} (x_j^{(i)} - \mu_{y^{(i)}j})^2}{\sum_{i=1}^{m} 1\{y^{(i)} \text{ is odd}\}}$$

Hence, the two expressions are identical

Similarly for $\hat{\Sigma}_2$ i.e.

$$\sigma_{j2}^2 = \frac{\sum_{i=1}^{m} 1\{y^{(i)} \text{ is even}\} (x_j^{(i)} - \mu_{y^{(i)}j})^2}{\sum_{i=1}^{m} 1\{y^{(i)} \text{ is even}\}}$$

Q.4. (a)

$\theta^T A \theta$:- Following derivation works for both when A is symmetric or not symmetric.

$$\frac{\partial \theta^T A \theta}{\partial \theta_j} = \frac{\partial \sum_{k,k'} \theta_k A_{kk'} \theta_{k'}}{\partial \theta_j}$$

... where $k = j$ or $k' = j$ will remain.

Terms with $k = j$ ...

$$\frac{\partial}{\partial \theta_j}\left[\sum_{k=j,\, k'\neq j} \theta_j A_{jk'} \theta_{k'} + \sum_{k\neq j,\, k'=j} \theta_k A_{kj} \theta_j + \sum_{k=j,\, k'=j} \theta_j A_{jj} \theta_j\right]$$

$$= \sum_{k'\neq j} A_{jk'} \theta_{k'} + \sum_{k\neq j} \theta_k A_{kj} + 2\theta_j A_{jj}$$

A^T:- transpose of A

$$= \sum_{k'} A_{jk'} \theta_{k'} + \sum_{k} \theta_k A_{kj} = \sum_{k'} A_{jk'} \theta_{k'} + \sum_{k} A^T_{jk} \theta_k$$

$$= (A\theta + A^T\theta)_j$$

$$\Rightarrow \nabla_\theta \theta^T A \theta = (A + A^T)\theta$$

$$\frac{\partial^2 \theta^T A \theta}{\partial \theta_j \partial \theta_l} = \frac{\partial}{\partial \theta_l}\left[\sum_{k'} A_{jk'} \theta_{k'} + \sum_{k} A^T_{jk} \theta_k\right]$$

$$= (A_{jl} + A^T_{jl})$$

$$\Rightarrow \nabla^2_\theta \theta^T A \theta = (A + A^T)$$

$f(\theta) = \theta^T A \theta + a^T \theta + b$

$\Rightarrow$ Newton's update:-

$$\theta^{(t+1)} \leftarrow \theta^{(t)} - H^{-1} \nabla_\theta f(\theta)$$

$$= \theta^{(t)} - (A+A^T)^{-1}\left[(A+A^T)\theta^{(t)} + a\right]$$

$$= \cancel{\theta^{(t)} - I\theta^{(t)}} - (A+A^T)^{-1}a$$

$$= -(A+A^T)^{-1}a \quad \text{(constant wrt } \theta)$$

$\Rightarrow$ Newton's method converges in single iteration

(b) GD:-

(i)

$$\theta^{(t+1)} \leftarrow \theta^{(t)} + \eta \cdot \frac{1}{m} \nabla_\theta LL(\theta)$$

$$[\theta^{(t+1)} - \theta^{(t)}] \leftarrow \eta \cdot \frac{1}{m} \nabla_\theta LL(\theta)$$

SGD:-

$$\theta^{(t+1)} \leftarrow \theta^{(t)} + \eta \cdot \frac{1}{\gamma} \nabla_\theta LL_b(\theta)$$

$$\Rightarrow \theta^{(t+1)} - \theta^{(t)} \leftarrow \eta \cdot \frac{1}{\gamma} \nabla_\theta LL_b(\theta)$$

$$E_{\{x^{(i)}, y^{(i)}\}_{i=1}^{\gamma}} \left[\theta^{(t+1)} - \theta^{(t)}\right] \leftarrow \eta \cdot \frac{1}{\gamma} E_{\{x^{(i)}, y^{(i)}\}_{i=1}^{\gamma}} \left[\nabla_\theta LL_b(\theta)\right]$$

$\forall i :- (x^{(i)}, y^{(i)}) \sim D$ $\qquad$ $\forall i : (x^{(i)}, y^{(i)}) \sim D$

D:- Training data distribution

RHS

$$= \eta \cdot \frac{1}{\gamma} E_{\{x^{(i)}, y^{(i)}\}_{i=1}^{m}} \nabla_\theta \left[\sum_{i=1}^{\gamma} LL_i(\theta)\right]$$

$\forall i : (x^{(i)}, y^{(i)}) \sim D$

where $LL_i(\theta)$ denotes the log-likelihood computed over the ith example in mini-batch

$$= \eta \cdot \frac{1}{\gamma} \nabla_\theta \, E_{\substack{\{x^{(i)}, y^{(i)}\}_{i=1}^{\gamma} \\ \forall i : (x^{(i)}, y^{(i)}) \sim D}} \sum_{i=1}^{\gamma} LL_i(\theta)$$

$$= \eta \cdot \frac{1}{\gamma} \nabla_\theta \sum_{i=1}^{\gamma} E_{(x^{(i)}, y^{(i)}) \sim D} [LL_i(\theta)]$$

$$= \eta \cdot \frac{1}{\gamma} \nabla_\theta \sum_{i=1}^{\gamma} \left[\frac{1}{m} \sum_{l=1}^{m} LL_l(\theta) \right]$$

$$= \eta \cdot \frac{1}{\cancel{\gamma}} \nabla_\theta \cancel{\gamma} \underbrace{\frac{1}{m} \sum_{l=1}^{m} LL_l(\theta)}_{LL(\theta)}$$

$$= \eta \cdot \nabla_\theta LL(\theta)$$

$\Rightarrow$ Same as GD update

(ii)

In SGD:-

$$\theta^{(t+1)} \leftarrow \theta^{(t)} + \eta \cdot \frac{1}{\gamma} \nabla_\theta LL_S(\theta)$$

$$\theta^{(t+1)} - \theta^{(t)} \leftarrow \eta \cdot \frac{1}{\gamma} \nabla_\theta LL_S(\theta)$$

$$Var(\theta^{(t+1)} - \theta^{(t)}) \leftarrow \eta^2 \frac{1}{\gamma^2} Var[\nabla_\theta LL_S(\theta)]$$

$$RHS = \eta^2 \frac{1}{\gamma^2} Var[\nabla_\theta LL_S(\theta)]$$

$$= \eta^2 \frac{1}{\gamma^2} Var\left[\nabla_\theta \sum_{i=1}^{\gamma} LL_i(\theta)\right]$$

$LL_i(\theta)$ $\Downarrow$ Log-likelihood for ith example

where the variance is computed wrt distribution $\{x^{(i)}, y^{(i)}\}_{i=1}^{\gamma}$, $(x^{(i)}, y^{(i)}) \sim D$ (training data dist.)

$$\eta^2 \frac{1}{\gamma^2} Var\left[\sum_{i=1}^{\gamma} \nabla_\theta LL_i(\theta)\right]$$

$$\cdots \frac{1}{\gamma^2} \cdots \sum_{i=1} \underbrace{\qquad\qquad}_{\gamma \text{ independent random variables}}$$

$$\Rightarrow \eta^2 \frac{1}{\gamma^2} \sum_{i=1}^{\gamma} Var\left[\nabla_\theta LL_i(\theta)\right]$$

But $Var[\nabla_\theta LL_i(\theta)] = Var[\nabla_\theta LL_1(\theta)]$

By definition (each of $\nabla_\theta LL_i(\theta)$ are i.i.d.)

$$\Rightarrow Var(\theta^{(t+1)} - \theta^{(t)}) = \eta^2 \frac{1}{\gamma^2} \gamma \; Var[\nabla_\theta LL_1(\theta)]$$

$$= \eta^2 \frac{1}{\gamma} \; Var[\nabla_\theta LL_1(\theta)]$$