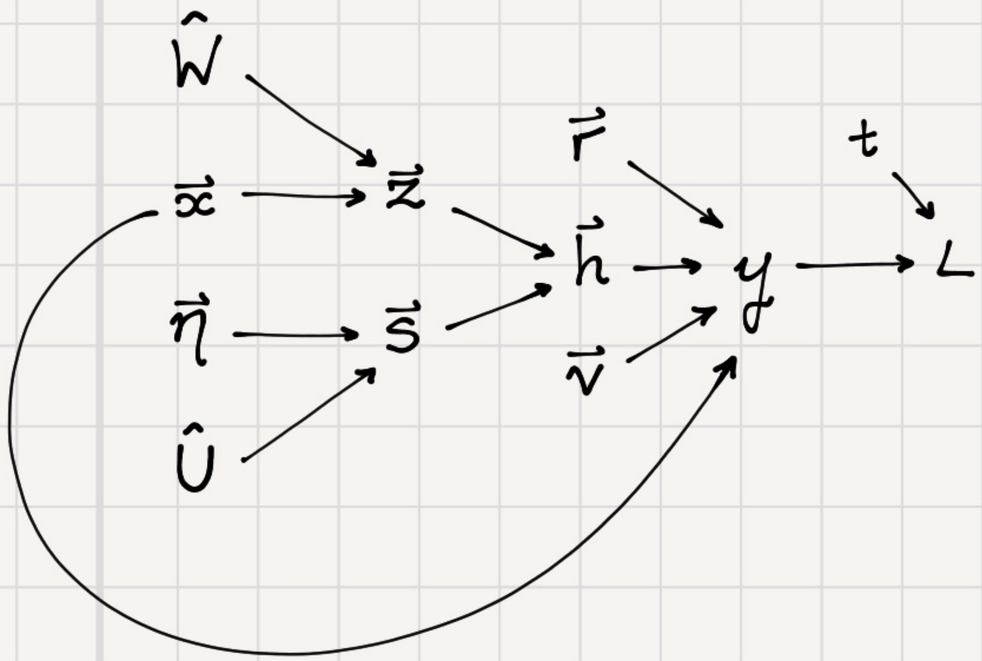
Homework 3

(1) Backprop

a)



b) $\frac{d}{dx}\sigma(x) = \frac{d}{dx}\left(\frac{1}{1+e^{-x}}\right) = \frac{d}{dx}\left[(1+e^{-x})^{-1}\right]$

$= -(1+e^{-x})^{-2}(-e^{-x})$

$= e^{-x}(1+e^{-x})^{-2}$

$= \frac{e^{-x}}{(1+e^{-x})^2} = \frac{1}{1+e^{-x}} \cdot \frac{e^{-x}}{1+e^{-x}}$

$= \sigma(x)(1-\sigma(x))$

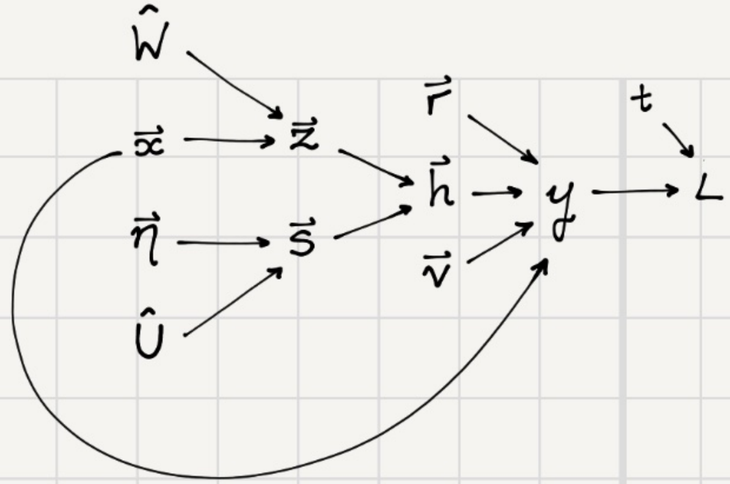
Rough work

$1-\sigma(x) = 1 - \frac{1}{1+e^{-x}}$

$= \frac{1+e^{-x}}{1+e^{-x}} - \frac{1}{1+e^{-x}}$

$= \frac{e^{-x}}{1+e^{-x}}$

Homework 3



(1) Backprop

c) Define $a = \vec{v}^T\vec{h} + \vec{r}^T\vec{x}$, so $\bar{a} = t - y$

$\bar{L} = 1$

$$\bar{v} = \frac{\partial L}{\partial v} = \frac{\partial L}{\partial y} \cdot \frac{\partial y}{\partial v}$$

$$= \frac{t-y}{y(1-y)} \cdot h\big(y(1-y)\big) = (t-y)\vec{h} = \bar{a}\vec{h}$$

$$\bar{r} = \frac{\partial L}{\partial r} = \frac{\partial L}{\partial y} \cdot \frac{\partial y}{\partial r}$$

$$= \frac{t-y}{y(1-y)} \cdot \big(y(1-y)\big)x = (t-y)\vec{x} = \bar{a}\vec{x}$$

Rough Work

$$\frac{\partial L}{\partial y} = \frac{\partial}{\partial y}\Big(t\log y + (1-t)\log(1-y)\Big)$$

$$= \left(\frac{t}{y} - \frac{1-t}{1-y}\right) = \frac{t-y}{y(1-y)}$$

$$\frac{\partial y}{\partial v} = \frac{\partial}{\partial v}\Big(\sigma(v^Th + r^Tx)\Big) = h\sigma'(v^Th + r^Tx)$$

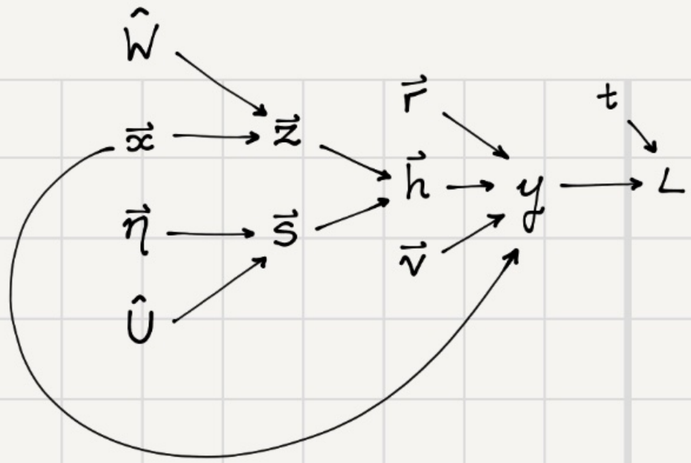
$$= h\Big[\sigma(v^Th + r^Tx)\big(1 - \sigma(v^Th + r^Tx)\big)\Big]$$

$$= h\big(y(1-y)\big)$$

$$\frac{\partial y}{\partial r} = \frac{\partial}{\partial r}\Big(\sigma(v^T\vec{h} + r^T\vec{x})\Big)$$

$$= \vec{x}\big(y(1-y)\big)$$

Homework 3



(1) Backprop

c) $$\overline{W} = \frac{\partial L}{\partial W} = \frac{\partial L}{\partial y} \cdot \frac{\partial y}{\partial h} \cdot \frac{\partial h}{\partial z} \cdot \frac{\partial z}{\partial W}$$

$$= \frac{t-y}{y(1-y)} \cdot y(1-y)\vec{v} \cdot \text{diag}(\vec{s}) \cdot \frac{\partial z}{\partial W}$$

$$= (t-y)(\vec{v} \odot \vec{s}) \cdot \frac{\partial (W\vec{x})}{\partial W}$$

Let $z \in \mathbb{R}^m$, where m is the number of rows of W. Then:

$$= (t-y)(\vec{v} \odot \vec{s})(\mathbf{1}_m \otimes \vec{x}^T)$$

$$= (t-y)(\vec{v} \odot \vec{s})\vec{x}^T$$

$$= \overline{z}\vec{x}^T$$

Rough work

$$\frac{\partial \vec{y}}{\partial \vec{h}} = \frac{\partial}{\partial \vec{h}}\left(\sigma\left(\vec{v}^T\vec{h} + \vec{r}^T\vec{x}\right)\right) = (y(1-y))\vec{v}$$

$$\frac{\partial \vec{h}}{\partial \vec{z}} = \text{diag}(\vec{s}). \quad \text{E.g. } \vec{z} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}, \vec{s} = \begin{bmatrix} s_1 \\ s_2 \end{bmatrix}$$

$$\vec{h} = \vec{z} \odot \vec{s} = \begin{bmatrix} z_1 s_1 \\ z_2 s_2 \end{bmatrix}$$

$$\frac{\partial \vec{h}}{\partial \vec{z}} = \begin{bmatrix} \frac{\partial h_1}{\partial z_1} & \frac{\partial h_1}{\partial z_2} \\ \frac{\partial h_2}{\partial z_1} & \frac{\partial h_2}{\partial z_2} \end{bmatrix} = \begin{bmatrix} s_1 & 0 \\ 0 & s_2 \end{bmatrix} = \text{diag}(\vec{s})$$

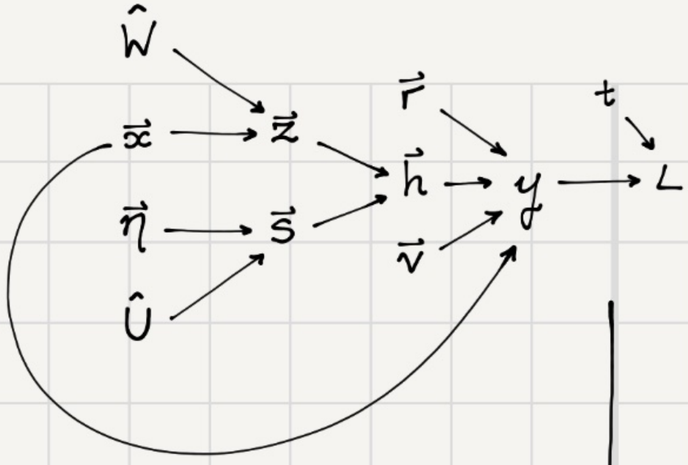
$$z = Wx \rightarrow \frac{\partial \vec{z}}{\partial W} = \frac{\partial (W\vec{x})}{\partial W}$$

$$z_i = \sum_j (W_{ij}x_j) \Rightarrow \frac{\partial z_i}{\partial W_{ij}} = x_j$$

and $$\frac{\partial z_i}{\partial W_{kj}} = \begin{cases} x_j & \text{if } k = i \\ 0 & \text{otherwise} \end{cases}$$

This lets us "flatten" the 3D tensor $\frac{\partial \vec{z}}{\partial W}$ into just $\vec{x}^T$ (row-wise repeated)

Homework 3



(1) Backprop

c) $\overline{U} = \frac{\partial L}{\partial U} = \frac{\partial L}{\partial y} \cdot \frac{\partial y}{\partial h} \cdot \frac{\partial h}{\partial s} \cdot \frac{\partial s}{\partial U}$

$= \frac{(t-y)}{y(1-y)} \cdot y(1-y)\vec{v} \cdot \text{diag}(\vec{z}) \cdot \vec{\eta}^T$

$= (t-y)(\vec{v} \odot \vec{z})\vec{\eta}^T = \vec{s}\vec{\eta}^T$

$\overline{\eta} = \frac{\partial L}{\partial \eta} = \frac{\partial L}{\partial y} \cdot \frac{\partial y}{\partial h} \cdot \frac{\partial h}{\partial s} \cdot \frac{\partial s}{\partial \eta}$

$= \frac{(t-y)}{y(1-y)} \cdot y(1-y)\vec{v} \cdot \text{diag}(\vec{z}) \cdot \frac{\partial (U\eta)}{\partial \eta}$

$= U^T(t-y) \cdot (\vec{v} \odot \vec{z}) = U^T\vec{s}$

skip connection

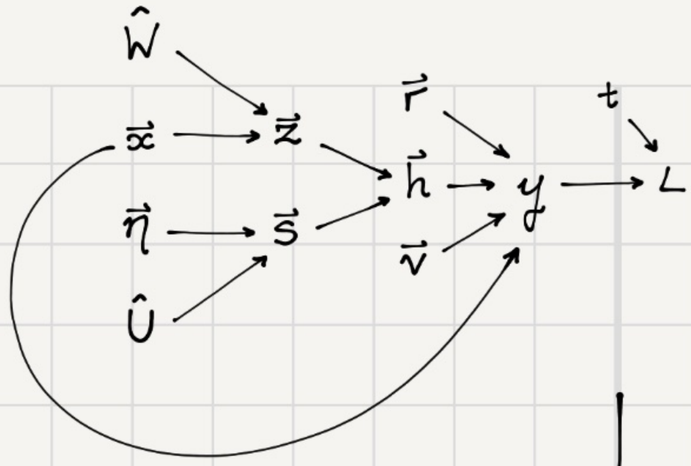
$\overline{x} = \frac{\partial L}{\partial x} = \frac{\partial L}{\partial y} \cdot \left(\frac{\partial y}{\partial h} \cdot \frac{\partial h}{\partial z} \cdot \frac{\partial z}{\partial x} + \frac{\partial y}{\partial x} \right)$

$= \frac{t-y}{y(1-y)} \cdot \left(y(1-y)\vec{v} \cdot \text{diag}(\vec{s}) \cdot W + y(1-y)\vec{r} \right)$

$= W^T(t-y)(\vec{v} \odot \vec{s}) + (t-y)\vec{r}$

$= W^T\vec{s} + \vec{a}\vec{r}$

Homework 3



(1) Backprop

c) To summarize:

· Seed: $\bar{L} = 1$

· Output node: $\bar{y} = \frac{t-y}{y(1-y)}$

· Pre-activation: $a = \vec{v}^T\vec{h} + \vec{r}^T\vec{x}$

$\bar{a} = t - y$

· Hidden layer: $\bar{h} = \bar{a}\vec{v}$

· Parameters:

$\bar{v} = \bar{a}\vec{h}$; $\bar{r} = \bar{a}\vec{x}$

$\bar{W} = \bar{z}\cdot\vec{x}^T$; $\bar{U} = \bar{s}\cdot\bar{\eta}^T$

$\bar{z} = \bar{a}(\vec{v}\odot\vec{s})$; $\bar{s} = \bar{a}(\vec{v}\odot\vec{z})$

$\bar{\eta} = U^T\bar{s}$

· Input: $\bar{x} = W^T\bar{z} + \bar{a}\vec{r}$

Homework 3

(2) Fitting a Naïve Bayes Model

a) To derive the maximum likelihood estimator (MLE) for the class-conditional pixel probabilities $\hat{\theta}$ and the prior $\vec{\pi}$, we take the derivative of the log-likelihood and set it to zero. To do this we must first find the likelihood function:

$$\mathcal{L}(\theta, \pi) = \prod_{i=1}^{N} p\left(t^{(i)}, x^{(i)} \mid \theta, \pi\right)$$

(since MNIST images $x^{(i)}$ are i.i.d)

$$= \prod_{i=1}^{N} p\left(t^{(i)} \mid \pi\right) \cdot p\left(x^{(i)} \mid t^{(i)}, \theta\right)$$

(using Bayes Rule & Chain Rule)

- We may express the prior as:

$$p\left(t^{(i)} \mid \pi\right) = \prod_{c=0}^{9} \pi_c^{t_c^{(i)}}$$

since this will simply give the probability π_c of the class $c \in \{0, 1, \ldots, 9\}$ we are currently looping over, since $t_c^{(i)} = 1$ for the true class c, and $t_k^{(i)} = 0 \quad \forall k \neq c$.

a) · Next, we may express the likelihood as

$$p(x^{(i)} \mid t^{(i)}, \theta) = \prod_{c=0}^{9} \prod_{j=0}^{783} \theta_{jc}^{x_j^{(i)} t_c^{(i)}} (1-\theta_{jc})^{(1-x_j^{(i)}) t_c^{(i)}}$$

because we assume that, conditioned on the class, whether any pixel is "on" or "off" independent of all other pixels (Naïve Bayes assumption), and only the terms corresponding to the true class contribute due to the one-hot structure of $t^{(i)}$.

· Putting our expressions for the prior and the likelihood together, we have:

$$\mathcal{L}(\theta, \pi) = \prod_{i=1}^{N} \prod_{c=0}^{9} \left[\pi_c \prod_{j=0}^{783} \theta_{jc}^{x_j^{(i)}} (1-\theta_{jc})^{(1-x_j^{(i)})} \right]^{t_c^{(i)}}$$

· The log-likelihood is therefore

$$\ell(\theta, \pi) = \log \mathcal{L}(\theta, \pi) = \sum_{i=1}^{N} \sum_{c=0}^{9} \left[t_c^{(i)} \log(\pi_c) + \sum_{j=0}^{783} t_c^{(i)} \log\left(\theta_{jc}^{x_j^{(i)}} (1-\theta_{jc})^{(1-x_j^{(i)})} \right) \right]$$

$$= \sum_{i=1}^{N} \sum_{c=0}^{9} \left\{ t_c^{(i)} \log(\pi_c) + \sum_{j=0}^{783} t_c^{(i)} \left[x_j^{(i)} \log(\theta_{jc}) + (1-x_j^{(i)}) \log(1-\theta_{jc}) \right] \right\}$$

a) (continued)

$$\ell(\theta,\pi) = \sum_{i=1}^{N}\sum_{c=0}^{9} t_c^{(i)} \log(\pi_c) + \sum_{i=1}^{N}\sum_{c=0}^{9} t_c^{(i)} \sum_{j=0}^{783} \left[x_j^{(i)} \log(\theta_{jc}) + (1-x_j^{(i)}) \log(1-\theta_{jc}) \right]$$

$$= \sum_{c=0}^{9} \left[\left(\sum_{i=1}^{N} t_c^{(i)} \right) \log(\pi_c) + \sum_{j=0}^{732} \left(\sum_{i=1}^{N} t_c^{(i)} x_j^{(i)} \log(\theta_{jc}) + \sum_{i=1}^{N} t_c^{(i)} (1-x_j^{(i)}) \log(1-\theta_{jc}) \right) \right]$$

· Define $N_c = \sum_{i=1}^{N} t_c^{(i)}$, $S_{jc}^{(1)} = \sum_{i=1}^{N} t_c^{(i)} x_j^{(i)}$, $S_{jc}^{(0)} = \sum_{i=1}^{N} t_c^{(i)} (1-x_j^{(i)})$. These are constant w.r.t π and θ. Then:

$$\ell(\theta,\pi) = \sum_{c=0}^{9} \left(N_c \log(\pi_c) + \sum_{j=0}^{732} \left(S_{jc}^{(1)} \log(\theta_{jc}) + S_{jc}^{(0)} \log(1-\theta_{jc}) \right) \right)$$

· Then we take $\partial\ell/\partial\theta_{jc}$ and $\partial\ell/\partial\pi_c$ and set both to zero:

$$\frac{\partial \ell}{\partial \theta_{jc}} = \sum_{c=0}^{9} \left(\sum_{j=0}^{732} \left(\frac{S_{jc}^{(1)}}{\theta_{jc}} + \frac{S_{jc}^{(0)}}{(1-\theta_{jc})} \right) \right) = 0 \Rightarrow \frac{S_{jc}^{(1)}}{\theta_{jc}} + \frac{S_{jc}^{(0)}}{(1-\theta_{jc})} = 0 \quad \forall j,c$$

$$\therefore \hat{\theta}_{jc} = \frac{S_{jc}^{(1)}}{S_{jc}^{(1)} + S_{jc}^{(0)}} = \frac{S_{jc}^{(1)}}{N_c}$$

a) . $\frac{\partial l}{\partial \pi_c} = \frac{\partial}{\partial \pi_c}\left(\sum_{c=0}^{9} N_c \log(\pi_c)\right)$. We replace $\pi_9 = 1 - \sum_{c=0}^{8} \pi_c$

. Then $l(\pi) = \sum_{c=0}^{8} N_c \log(\pi_c) + N_9 \log\left(1 - \sum_{c=0}^{8} \pi_c\right)$

$$\Rightarrow \frac{\partial l}{\partial \pi_c} = \frac{N_c}{\pi_c} - \frac{N_9}{1 - \sum_{c=1}^{8} \pi_c} = 0 \quad \forall c \quad \Rightarrow \frac{N_c}{\pi_c} = \frac{N_9}{\pi_9} \Rightarrow \frac{\pi_c}{\pi_9} = \frac{N_c}{N_9}$$

. Then $\pi_c = \pi_9 \frac{N_c}{N_9} \Rightarrow \sum_{c=0}^{8} \pi_c = 1 - \pi_9 = \pi_9 \sum_{c=0}^{8} \frac{N_c}{N_9}$

. Thus $N_9 (1 - \pi_9) = \pi_9 \sum_{c=0}^{8} N_c \Rightarrow N_9 - \pi_9 N_9 = \pi_9 \sum_{c=0}^{8} \frac{N_c}{N_9}$

. $N_9 = \pi_9 \left(N_9 + \sum_{c=0}^{8} N_c\right) = \pi_9 \sum_{c=0}^{9} N_c \Rightarrow \pi_9 = \frac{N_9}{\sum_{c=0}^{9} N_c} = \frac{N_9}{N}$

$$\therefore \hat{\pi}_c = \frac{N_c}{\sum_{c=0}^{9} N_c} = \frac{N_c}{N}$$

(2) b) $$p(t|x,\theta,\pi) = \frac{p(t,x|\theta,\pi)}{p(x|\theta,\pi)} = \frac{p(t|\pi)\cdot p(x|t,\theta)}{p(x|\theta,\pi)}$$ (using Bayes Rule & Chain Rule)

$$\Rightarrow \log p(t|x,\theta,\pi) = \log\{p(t|\pi)\cdot p(x|t,\theta)\} - \log p(x|\theta,\pi)$$

· We may express the prior as:

$$p(t|\pi) = \prod_{c=0}^{9} \pi_c^{t_c},$$

since this will simply give the probability π_c of the class $c \in \{0,1,\dots,9\}$ we are currently looping over, since $t_c = 1$ for the true class c, and $t_k = 0 \ \forall k \neq c$. (By t_c I mean the c^{th} component of t).

· Next, we may express the likelihood as

$$p(x|t,\theta) = \prod_{c=0}^{9} \prod_{j=0}^{783} \theta_{jc}^{x_j t_c} (1-\theta_{jc})^{(1-x_j)t_c}$$

because we assume that, conditioned on the class, whether any pixel is "on" or "off" independent of all other pixels (Naïve Bayes assumption), and only the terms corresponding to the true class contribute due to the one-hot structure of t.

(2) b) The normalizer is:

$$p(x|\hat{\theta},\hat{\pi}) = \sum_{k=0}^{9} \pi_k \prod_{j=0}^{783} \theta_{jk}^{x_j} (1-\theta_{jk})^{(1-x_j)}$$

Putting our expressions for the prior, likelihood, and normalizer together, and subbing in the MLEs $\hat{\theta}$ and $\hat{\pi}$:

$$p(t|x,\hat{\theta},\hat{\pi}) = \prod_{c=0}^{9} \left[\hat{\pi}_c \prod_{j=0}^{783} \hat{\theta}_{jc}^{x_j} (1-\hat{\theta}_{jc})^{(1-x_j)} \right]^{t_c} - \sum_{k=0}^{9} \hat{\pi}_k \prod_{j=0}^{783} \hat{\theta}_{jk}^{x_j} (1-\hat{\theta}_{jk})^{(1-x_j)}$$

Taking the logarithm:

$$\log p(t|x,\hat{\theta},\hat{\pi}) = \sum_{c=0}^{9} t_c \left\{ \log(\hat{\pi}_c) + \sum_{j=0}^{783} \left[x_j \log(\hat{\theta}_{jc}) + (1-x_j) \log(1-\hat{\theta}_{jc}) \right] \right\}$$

$$- \log\left(\sum_{k=0}^{9} \pi_k \prod_{j=0}^{783} \hat{\theta}_{jk}^{x_j} (1-\hat{\theta}_{jk})^{(1-x_j)} \right)$$

Homework 3

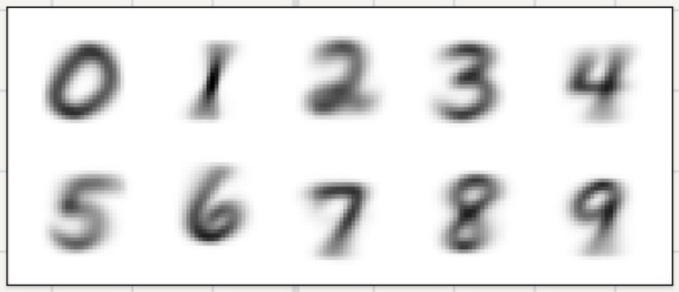
(2) Fitting a Naïve Bayes Model

c) We attempt to compute the average log-likelihood per data point using:

$$\frac{1}{N}\sum_{i=1}^{N} \log p\left(t^{(i)} \mid x^{(i)}, \hat{\theta}, \hat{\pi}\right)$$

However this quantity isn't computable in the discriminative (logistic reg.) setting, because we don't have access to the full joint distribution $p(t,x)$. Logistic regression only models the conditional dist. $p(t|x)$, not the marginal over x, which is required to compute the full log-likelihood under a generative model.

d) MLE estimator:



Homework 3

(2) Fitting a Naïve Bayes Model

e) For class c, suppose we have N_c images. Let N_{jc} be the number of images where pixel j is on.

- The likelihood of the pixel values under a Bernoulli model is

$$\mathcal{L}(\theta_{jc}) = \theta_{jc}^{N_{jc}} (1-\theta_{jc})^{N_c - N_{jc}}$$

- We assume the prior follows a Beta distribution

$$\theta_{jc} \sim \text{Beta}(\alpha, \beta) \Rightarrow p(\theta_{jc}) \propto \theta_{jc}^{\alpha-1} (1-\theta_{jc})^{\beta-1}$$

- The MAP estimator maximizes the posterior:

$$p(\theta_{jc} | \text{data}) \propto \mathcal{L}(\theta_{jc}) p(\theta_{jc}) = \theta_{jc}^{N_{jc}+\alpha-1} (1-\theta_{jc})^{N_c - N_{jc} + \beta - 1}$$

Homework 3

(2) Fitting a Naïve Bayes Model

e) (continued) · We then take the log and maximize:

$$\log p(\theta_{jc} | data) = (N_{jc} + \alpha - 1)\log(\theta_{jc}) + (N_c - N_{jc} + \beta - 1)\log(1 - \theta_{jc})$$

· Taking the derivative w.r.t θ_{jc} and setting to 0:

$$\frac{d\log(p)}{d\theta_{jc}} = \frac{N_{jc} + \alpha - 1}{\theta_{jc}} - \frac{N_c - N_{jc} + \beta - 1}{1 - \theta_{jc}} = 0$$

$$\Rightarrow (N_{jc} + \alpha - 1)(1 - \theta_{jc}) = (N_c - N_{jc} + \beta - 1)\,\theta_{jc}$$

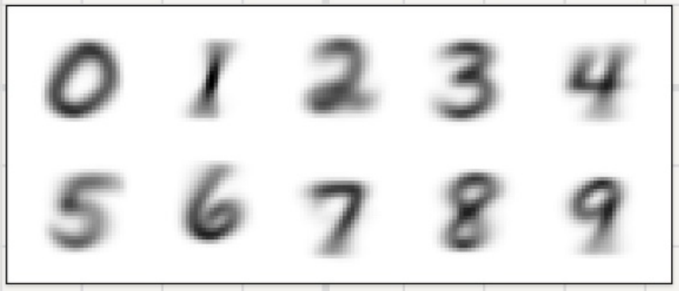
$$\Rightarrow \boxed{\hat{\theta}_{jc}^{MAP} = \frac{N_{jc} + \alpha - 1}{N_c + \alpha + \beta - 2}} \quad \cdot \text{ For } \alpha = \beta = 3, \text{ this is: } \hat{\theta}_{jc}^{MAP} = \frac{N_{jc} + 2}{N_c + 4}$$

Homework 3

(2) Fitting a Naïve Bayes Model

f) Average log-likelihood for MLE: -173 | Training Accuracy for MLE: 0.835
Average log-likelihood for MAP: -173 | Training Accuracy for MAP: 0.816

g) MAP estimator:



h) Advantage: Naïve Bayes is a fast and simple algorithm. Training just needs one pass through the data (i.e when estimating θ_{jc} and π_c). That said, the conditional independence assumption is unrealistic for these images. Pixels of a handwritten digit are highly correlated, but Naïve Bayes treats them as independent given the class/digit. That mismatch limits accuracy. We furthermore must rely on smoothing tricks to avoid log(0) issues

Homework 3

(3) Logistic Regression with Gaussian Prior

a) • Under logistic regression, for a single pair $(\vec{x}^{(i)}, y^{(i)}) \in D$, we have

$$P(y^{(i)}=1 \mid \vec{x}^{(i)}, \vec{\theta}) = \sigma(\vec{x}^{(i)T}\vec{\theta}), \text{ and } P(y^{(i)}=0 \mid \vec{x}^{(i)}, \vec{\theta}) = 1-\sigma(\vec{x}^{(i)T}\vec{\theta})$$

where $\sigma(z) = \frac{1}{1+e^{-z}}$

We may thus compactly write:

$$P(y^{(i)} \mid \vec{x}^{(i)}, \vec{\theta}) = \left(\sigma(\vec{x}^{(i)T}\vec{\theta})\right)^{y^{(i)}} \left(1-\sigma(\vec{x}^{(i)T}\vec{\theta})\right)^{(1-y^{(i)})}$$

• Thus, for the entire dataset:

$$P(\vec{y} \mid \hat{X}, \vec{\theta}) = \prod_{i=1}^{N} P(y^{(i)} \mid \vec{x}^{(i)}, \vec{\theta}) = \prod_{i=1}^{N} \left(\sigma(\vec{x}^{(i)T}\vec{\theta})\right)^{y^{(i)}} \left(1-\sigma(\vec{x}^{(i)T}\vec{\theta})\right)^{(1-y^{(i)})}$$

$$\Rightarrow \log P(\vec{y} \mid \hat{X}, \vec{\theta}) = \sum_{i=1}^{N} \left\{ y^{(i)} \log\left[\sigma(\vec{x}^{(i)T}\vec{\theta})\right] + (1-y^{(i)}) \log\left[1-\sigma(\vec{x}^{(i)T}\vec{\theta})\right] \right\}$$

Homework 3

(3) Logistic Regression with Gaussian Prior

a) $$\frac{d}{d\theta}\left(\log p(\vec{y} \mid \hat{X}, \vec{\theta})\right) = \sum_{i=1}^{N} \left(\frac{y^{(i)}\vec{x}^{(i)}}{\sigma(\vec{x}^{(i)T}\vec{\theta})} \cdot \sigma(\vec{x}^{(i)T}\vec{\theta})\left(1-\sigma(\vec{x}^{(i)T}\vec{\theta})\right) - \frac{(1-y^{(i)})(\vec{x}^{(i)})}{1-\sigma(\vec{x}^{(i)T}\vec{\theta})} \sigma(\vec{x}^{(i)T}\vec{\theta})\left(1-\sigma(\vec{x}^{(i)T}\vec{\theta})\right) \right)$$

$$= \sum_{i=1}^{N} \left(y^{(i)}\left(1-\sigma(\vec{x}^{(i)T}\vec{\theta})\right) - (1-y^{(i)})\sigma(\vec{x}^{(i)T}\vec{\theta}) \right) \vec{x}^{(i)}$$

$$= \sum_{i=1}^{N} \left(y^{(i)} - \sigma(\vec{x}^{(i)T}\vec{\theta}) \right) \vec{x}^{(i)} = 0$$

There is no closed form solution that sets this gradient to zero; we can't explicitly solve for θ, since it is tied to the exponential $e^{-x^T\theta}$. Thus, we would need to minimize the negative log-likelihood using an iterative approach such as gradient descent:

$$\theta \leftarrow \theta + \eta \sum_{i=1}^{N} \left(y^{(i)} - \sigma(\vec{x}^{(i)T}\vec{\theta}) \right) \vec{x}^{(i)}$$

Homework 3

(3) Logistic Regression with Gaussian Prior

b) $$\hat{\theta}_{MAP} = \arg\max_{\theta} \left\{ p(\vec{\theta} \mid X, \vec{y}) \right\} = \arg\max_{\theta} \left\{ p(\vec{\theta})\, p(X, \vec{y} \mid \vec{\theta}) \right\}$$

$$= \arg\max_{\theta} \left\{ \log p(\vec{\theta}) + \log p(X, \vec{y} \mid \vec{\theta}) \right\}$$

- Assuming a Gaussian Prior $p(\vec{\theta}) \sim \mathcal{N}(0, \sigma^2 I)$, we have

$$\log p(\vec{\theta}) = \log \left(\frac{1}{(2\pi\sigma^2)^{d/2}} e^{-\frac{1}{2}\left(\frac{\|\vec{\theta}\|}{\sigma}\right)^2} \right) = \log\left((2\pi\sigma^2)^{-\frac{d}{2}} \right) + \log\left(e^{-\frac{1}{2}\left(\frac{\|\vec{\theta}\|}{\sigma}\right)^2} \right)$$

$$= -\frac{d}{2}\log(2\pi\sigma^2) - \frac{1}{2}\left(\frac{\|\vec{\theta}\|}{\sigma}\right)^2 \log(e) = \underbrace{-\frac{d}{2}\log(2\pi\sigma^2)}_{\text{constant w.r.t } \theta} - \frac{1}{2}\left(\frac{\|\vec{\theta}\|}{\sigma}\right)^2$$

$$p(X \mid \vec{\theta}) = p(X) \Rightarrow p(X, \vec{y} \mid \vec{\theta}) = p(y \mid X, \vec{\theta})\, p(X)$$

$$\Rightarrow \log p(X, \vec{y} \mid \theta) = \underbrace{\log p(\vec{y} \mid X, \vec{\theta})}_{\text{result from (a)}} + \underbrace{\log p(X)}_{\text{constant w.r.t } \theta}$$

Homework 3

(3) Logistic Regression with Gaussian Prior

b) The log-likelihood is thus: $\ell(\vec{\theta}) = \log p(\vec{\theta}) + \log p(\vec{y} \mid \hat{X}, \vec{\theta})$

For our purposes we may ignore the constants w.r.t θ, since they will not contribute to the derivative, so our log-likelihood is:

$$\ell(\vec{\theta}) = -\frac{1}{2}\left(\frac{\|\vec{\theta}\|}{\sigma}\right)^2 + \sum_{i=1}^{N} \left\{ y^{(i)} \log\left[\sigma(\vec{x}^{(i)T}\vec{\theta})\right] + (1-y^{(i)}) \log\left[1-\sigma(\vec{x}^{(i)T}\vec{\theta})\right] \right\}$$

Homework 3

(4) Gaussian Discriminant Analysis

a & b) We classify each point $\vec{x}^{(i)}$ by choosing the class with the highest posterior probability:

$$y^{(i)} = \arg\max_k \left\{ \log p\left(y^{(i)} = k \mid \vec{x}^{(i)}\right) \right\}$$

Using the full covariance model, we achieve:

- Average log-likelihood (train): -0.125
- Average log-likelihood (test) : -0.197

- Training accuracy: 98.1%
- Test accuracy: 97.3%

c) Using a diagonal approximation to the covariance matrix (i.e, assuming all pixels are independent), we obtain

- Average log-likelihood (train): -1.23
- Average log-likelihood (test) : -1.29

Training accuracy: 85.0%
Test accuracy: 84.0%

Homework 3

(4) Gaussian Discriminant Analysis

c) The full covariance model captures correlations between different pixels, which allows it to better model the true distribution of the image data. This leads to higher log-likelihoods and accuracies, as seen in parts (a) and (b). E.g., it achieves 97.8% test accuracy vs 84% using the diagonal model

In contrast, the diagonal model assumes the pixels are independent given the class, which is a much stronger assumption as well as an unrealistic one, especially for images of handwritten digits where neighboring pixels are highly correlated. This independence assumption results in poorer modeling of the data.

That said, the diagonal model is computationally cheaper since estimating and inverting diagonal matrices is much faster and numerically stable compared to full matrices, so for large amounts of data it may be the preferred approach.