Homework 2

Problem 1

(a) $f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

$$L(\mu,\sigma^2) = \prod_{i=1}^{n} f(x_i;\mu,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$\ln L(\mu,\sigma^2) = \ln\left(\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right) = \sum_{i=1}^{n} \ln\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right)$$

$$= \sum_{i=1}^{n}\left[-\ln(\sigma) - \frac{1}{2}\ln(2\pi) - \frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

$$= -n\ln(\sigma) - \frac{n}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$\frac{\partial}{\partial\mu}\ln L(\mu,\sigma^2) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i - n\mu = 0 \Rightarrow \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\frac{\partial}{\partial\sigma^2}\ln L(\mu,\sigma^2) = -n\cdot\frac{1}{\sigma}\cdot\frac{1}{2\sigma} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 = 0$$

$$\Rightarrow -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 = 0 \Rightarrow \hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

Thus, $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$

(b) $f(x;\lambda) = \lambda e^{-\lambda x}$, where $x \geq 0$

$$L(\lambda) = \prod_{i=1}^{n} f(x_i;\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda\sum_{i=1}^{n} x_i}$$

$$\ln L(\lambda) = \ln\left(\lambda^n e^{-\lambda\sum_{i=1}^{n} x_i}\right) = n\ln(\lambda) - \lambda\sum_{i=1}^{n} x_i$$

$$\frac{\partial}{\partial\lambda}\ln L(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0 \Rightarrow \hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i}$$

Thus, $\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i}$

(c) $f(x;p) = p(1-p)^x$, where $x \geq 0$

$$L(p) = \prod_{i=1}^{n} f(x_i;p) = \prod_{i=1}^{n} p(1-p)^{x_i} = p^n(1-p)^{\sum_{i=1}^{n} x_i}$$

$$\ln L(p) = \ln\left(p^n(1-p)^{\sum_{i=1}^{n} x_i}\right) = n\ln(p) + \sum_{i=1}^{n} x_i\ln(1-p)$$

$$\frac{\partial}{\partial p}\ln L(p) = \frac{n}{p} - \frac{1}{1-p}\sum_{i=1}^{n} x_i = 0 \Rightarrow n(1-p) - p\sum_{i=1}^{n} x_i = 0$$

$$\Rightarrow n - (n + \sum_{i=1}^{n} x_i)p = 0 \Rightarrow \hat{p} = \frac{1}{1 + \frac{1}{n}\sum_{i=1}^{n} x_i}$$

Thus, $\hat{p} = \frac{1}{1 + \frac{1}{n}\sum_{i=1}^{n} x_i}$

Problem 2

(a) $p(y|\vec{x},\vec{w}) = \frac{1}{1+\exp(-y\vec{w}^T\vec{x})}$ $\quad p(y|\vec{x},\vec{w}) = Ber(y|sigm(\vec{w}^T\vec{x}))$

Since $Ber(y|p) = p^y(1-p)^{1-y}$,

then $Ber(y|sigm(\vec{w}^T\vec{x})) = sigm(\vec{w}^T\vec{x})^y(1-sigm(\vec{w}^T\vec{x}))^{1-y}$

$$= \left(\frac{1}{1+\exp(-\vec{w}^T\vec{x})}\right)^y\left(1-\frac{1}{1+\exp(-\vec{w}^T\vec{x})}\right)^{1-y}$$

$$= \frac{1}{(1+\exp(-\vec{w}^T\vec{x}))^y}\,\frac{\exp(-\vec{w}^T\vec{x})^{1-y}}{(1+\exp(-\vec{w}^T\vec{x}))^{1-y}}$$

$$= \frac{\exp(-\vec{w}^T\vec{x}(1-y))}{1+\exp(-\vec{w}^T\vec{x})}$$

If $y=1$, $Ber(y|sigm(\vec{w}^T\vec{x})) = \frac{1}{1+\exp(-\vec{w}^T\vec{x})}$

If $y=0$, $Ber(y|sigm(\vec{w}^T\vec{x})) = \frac{1}{1+\exp(\vec{w}^T\vec{x})}$

If $y=1$, $\frac{1}{1+\exp(-y\vec{w}^T\vec{x})} = \frac{1}{1+\exp(-\vec{w}^T\vec{x})}$

If $y=-1$, $\frac{1}{1+\exp(-y\vec{w}^T\vec{x})} = \frac{1}{1+\exp(-\vec{w}^T\vec{x})}$

Thus, Eq1 and Eq2 are equivalent representations for the case of binary classification

(b) $\hat{\vec{w}}_{MAP} = \text{argmax}\, p(\vec{w}|X,\vec{y})$

According to Baye's theorem, $p(\vec{w}|X,\vec{y}) \propto p(\vec{y}|X,\vec{w})\,p(\vec{w})$

$$p(\vec{y}|X,\vec{w}) = \prod_{i=1}^{n}\frac{1}{1+\exp(-y_i\vec{w}^T\vec{x})}$$

Since $\vec{w} \sim N(0,\sigma^2 I)$, $p(\vec{w}) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-\vec{w}^T\vec{w}}{2\sigma^2}\right)$

$$\ln p(\vec{w}|X,\vec{y}) \propto \ln p(\vec{y}|X,\vec{w}) + \ln p(\vec{w})$$

$$\ln p(\vec{y}|X,\vec{w}) = \ln \prod_{i=1}^{n}\frac{1}{1+\exp(-y_i\vec{w}^T\vec{x})} = -\sum_{i=1}^{n}\ln(1+\exp(-y_i\vec{w}^T\vec{x}))$$

$$\ln p(\vec{w}) = \ln\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-\vec{w}^T\vec{w}}{2\sigma^2}\right) = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{\vec{w}^T\vec{w}}{2\sigma^2}$$

$$nlp = \sum_{i=1}^{n}\ln(1+\exp(-y_i\vec{w}^T\vec{x})) + \frac{\vec{w}^T\vec{w}}{2\sigma^2}$$

(c) Since $nlp = \sum_{i=1}^{n} \ln(1+\exp(-y_i \vec{w}^T \vec{x})) + \frac{\vec{w}^T \vec{w}}{2\sigma^2}$

$$\min_{\vec{w}} \sum_{i=1}^{n} \ln(1+\exp(-y_i \vec{w}^T \vec{x})) + \frac{\vec{w}^T \vec{w}}{2\sigma^2}$$

Logistic loss: $L(\vec{w}) = \sum_{i=1}^{n} \ln(1+\exp(-y_i \vec{w}^T \vec{x}))$

L2 regularization: $R(\vec{w}) = \lambda \|\vec{w}\|^2 = \frac{1}{2\sigma^2} \|\vec{w}\|^2$, where $\lambda = \frac{1}{2\sigma^2}$

(d) since $p(y|\vec{x},\vec{w}) = \frac{1}{1+\exp(-y\vec{w}^T\vec{x})}$

$$p(y=1|\vec{x},\vec{w}) = \frac{1}{1+\exp(-\vec{w}^T\vec{x})} = \frac{\exp(\vec{w}^T\vec{x})}{1+\exp(\vec{w}^T\vec{x})}$$

$$p(y=-1|\vec{x},\vec{w}) = \frac{1}{1+\exp(\vec{w}^T\vec{x})}$$

$$\log \frac{p(y=1|\vec{x},\vec{w})}{p(y=-1|\vec{x},\vec{w})} = \log \frac{(1+\exp(\vec{w}^T\vec{x}))\exp(\vec{w}^T\vec{x})}{1+\exp(\vec{w}^T\vec{x})}$$

$$= \log \exp(\vec{w}^T\vec{x}) = \vec{w}^T\vec{x}$$

log odds: $\log \frac{p(y=1|\vec{x},\vec{w})}{p(y=-1|\vec{x},\vec{w})} = \vec{w}^T\vec{x}$

Each weight w_i represents the influence of the feature x_i on the prediction. Since, $p(y=1|\vec{x},\vec{w}) = \frac{1}{1+\exp(-\vec{w}^T\vec{x})}$

If $w_i > 0$, increasing x_i raises $p(y=1|\vec{x},\vec{w})$, if $w_i < 0$, increasing x_i lowers $p(y=1|\vec{x},\vec{w})$. Log odds change linearly with x_i, but probability changes non-linear due to the sigmoid function

Problem 3

(a) $P(X|Y) = P(X_1 = x_1, X_2 = x_2 | Y = y)$

$(X_1, X_2) \in \{(0,0), (0,1), (1,0), (1,1)\}$, $Y \in \{-1, 1\}$

$p(X_1=0, X_2=0|Y=1)$, $p(X_1=0, X_2=1|Y=1)$

$p(X_1=1, X_2=0|Y=1)$, $p(X_1=1, X_2=1|Y=1)$

$p(X_1=0, X_2=0|Y=-1)$, $p(X_1=0, X_2=1|Y=-1)$

$p(X_1=1, X_2=0|Y=-1)$, $p(X_1=1, X_2=1|Y=-1)$

$P(Y=1)$, $P(Y=-1)$

$2^2 \times 2 + 2 = 10$

Thus, we need to estimate 10 parameters.

(b) $2^{100} \times 2 + 2 = 2^{101} + 2$

Thus, we need to estimate $2^{101}+2$ parameters

(c) In problem (b), we observed that if we do not assume independence among the features, the number of parameters grows dramatically ($2^{d+1}+2$). This growth makes it computationally infeasible for large d, such as $d = 100$. But, with the naive assumption, the number of parameters reduces from $2^{d+1}+2$ to $4d+2$. For $d = 100$, this results in 402 parameters, which is computationally feasible.

(d) $h(\vec{x}) = \arg\max_c p(y=c) \prod_{\alpha=1}^{d} p(x_\alpha | y=c) = \text{sign}(\vec{w}_{BER}\vec{x} + b_{BER})$

Using Bayes' Theorem $p(y=c|x) \propto p(y=c) \prod_{\alpha=1}^{d} p(x_\alpha | y=c)$

For binary classification ($y \in \{-1,+1\}$), we compare

$$\ln p(y=1|x) - \ln p(y=-1|x)$$

since $\ln\left(p(y=c)\prod_{\alpha=1}^{d} p(x_\alpha|y=c)\right) = \ln p(y=c) + \sum_{\alpha=1}^{d} \ln p(x_\alpha|y=c)$

$\hat{\pi}_{+1} = p(y=+1)$, $\hat{\pi}_{-1} = p(y=-1)$

$\hat{\theta}_{\alpha,+1} = p(x_\alpha=1|y=+1)$, $\hat{\theta}_{\alpha,-1} = p(x_\alpha=1|y=-1)$

Since it is the Bernoulli naive Bayes classifier, then

$$p(x_\alpha|y) = \theta_{\alpha,y}^{x_\alpha}(1-\theta_{\alpha,y})^{1-x_\alpha}$$

Then, we have

$$\ln p(y=+1) + \sum_{\alpha=1}^{d} \ln p(x_\alpha|y=+1) - \ln p(y=-1) - \sum_{\alpha=1}^{d} \ln p(x_\alpha|y=-1)$$

$$= \left(\ln p(y=+1) - \ln p(y=-1)\right) + \sum_{\alpha=1}^{d}\left(\ln p(x_\alpha|y=+1) - \ln p(x_\alpha|y=-1)\right)$$

$$= \ln\frac{\hat{\pi}_{+1}}{\hat{\pi}_{-1}} + \sum_{\alpha=1}^{d} \ln\frac{\hat{\theta}_{\alpha+1}^{x_\alpha}(1-\hat{\theta}_{\alpha,+1})^{1-x_\alpha}}{\hat{\theta}_{\alpha,-1}^{x_\alpha}(1-\hat{\theta}_{\alpha,-1})^{1-x_\alpha}}$$

$$= \ln\frac{\hat{\pi}_{+1}}{\hat{\pi}_{-1}} + \sum_{\alpha=1}^{d}\left(x_\alpha \ln\frac{\hat{\theta}_{\alpha,+1}}{\hat{\theta}_{\alpha,-1}} + (1-x_\alpha)\ln\frac{1-\hat{\theta}_{\alpha,+1}}{1-\hat{\theta}_{\alpha,-1}}\right)$$

$$= \ln\frac{\hat{\pi}_{+1}}{\hat{\pi}_{-1}} + \sum_{\alpha=1}^{d}\left(x_\alpha \ln\frac{\hat{\theta}_{\alpha,+1}(1-\hat{\theta}_{\alpha,-1})}{\hat{\theta}_{\alpha,-1}(1-\hat{\theta}_{\alpha,+1})}\right) + \sum_{\alpha=1}^{d}\ln\frac{1-\hat{\theta}_{\alpha,+1}}{1-\hat{\theta}_{\alpha,-1}}$$

$$= \sum_{\alpha=1}^{d} x_\alpha w_{\alpha,BER} + b_{BER}$$

Thus, $\vec{w}_{\alpha,BER} = \left(\ln\frac{\hat{\theta}_{1,+1}(1-\hat{\theta}_{1,-1})}{\hat{\theta}_{1,-1}(1-\hat{\theta}_{1,+1})}, \ldots, \ln\frac{\hat{\theta}_{d,+1}(1-\hat{\theta}_{d,-1})}{\hat{\theta}_{d,-1}(1-\hat{\theta}_{d,+1})}\right)$

$$b_{BER} = \ln\frac{\hat{\pi}_{+1}}{\hat{\pi}_{-1}} + \sum_{\alpha=1}^{d}\ln\frac{1-\hat{\theta}_{\alpha,+1}}{1-\hat{\theta}_{\alpha,-1}}$$

Problem 4

(a) $P(Y=1)$, $P(Y=-1) = 1 - P(Y=1)$

Since $X_\alpha | Y \sim N(\mu_{\alpha,c}, \sigma^2_{\alpha,c})$

$$P(X_\alpha | Y) = \frac{1}{\sqrt{2\pi\sigma^2_{\alpha,c}}} \exp\left(-\frac{(x-\mu_{\alpha,c})^2}{2\sigma^2_{\alpha,c}}\right)$$

$$\hat{\mu}_{\alpha,c} = \frac{1}{N_c} \sum_{i:y_i=c} x_{\alpha,i}$$

$$\hat{\sigma}^2_{\alpha,c} = \frac{1}{N_c} \sum_{i:y_i=c} (x_{ai} - \hat{\mu}_{ac})^2$$

Suppose there are d features, and each feature X_α has $\mu_{\alpha,c}$ and $\sigma_{\alpha,c}$ in each category

For binary classification problem, the total number of parameters for Gaussian naive Bayes classifier is

$2d + 2d + 2 = 4d + 2$

(b) $P(Y=c | X_\alpha = x) \propto P(X_\alpha = x | Y=c) P(Y=c)$

$X_\alpha | Y \sim N(\mu_{\alpha,c}, \sigma^2_\alpha)$

$$P(X_\alpha = x | Y=y) = \frac{1}{\sqrt{2\pi\sigma^2_\alpha}} \exp\left(-\frac{(x-\mu_{\alpha c})^2}{2\sigma^2_\alpha}\right)$$

$$P(X=x | Y=y) = \prod_{\alpha=1}^{d} P(X_\alpha = x | Y=y)$$

$$\ln \frac{P(Y=+1|X)}{P(Y=-1|X)}$$

$$= \ln \frac{P(X|Y=1)P(Y=1)}{P(X|Y=-1)P(Y=-1)}$$

$$= \sum_{\alpha=1}^{d} \left[-\frac{(x_\alpha - \mu_{\alpha,+1})^2}{2\sigma^2_\alpha} + \frac{(x_\alpha - \mu_{\alpha,-1})^2}{2\sigma^2_\alpha} \right] + \ln \frac{\hat{\pi}_{+1}}{\hat{\pi}_{-1}}$$

$$= \sum_{\alpha=1}^{d} \left[-\frac{x_\alpha^2 - 2x_\alpha\mu_{\alpha,+1} + \mu_{\alpha,+1}^2}{2\sigma^2_\alpha} + \frac{x_\alpha^2 - 2x_\alpha\mu_{\alpha,-1} + \mu_{\alpha,-1}^2}{2\sigma^2_\alpha} \right] + \ln \frac{\hat{\pi}_{+1}}{\hat{\pi}_{-1}}$$

$$= \sum_{\alpha=1}^{d} \left(x_\alpha \frac{\mu_{\alpha,+1} - \mu_{\alpha,-1}}{\sigma^2_\alpha} \right) + \sum_{\alpha=1}^{d} \frac{\mu^2_{\alpha,-1} - \mu^2_{\alpha,+1}}{2\sigma^2_\alpha} + \ln \frac{\hat{\pi}_{+1}}{\hat{\pi}_{-1}}$$

$$\ln \frac{P(Y=1|X)}{P(Y=-1|X)} = w_0 + \sum_{\alpha=1}^{d} w_\alpha x_\alpha$$

$$w_\alpha = \frac{\mu_{\alpha,+1} - \mu_{\alpha,-1}}{\sigma^2_\alpha} \quad , \quad w_0 = \sum_{\alpha=1}^{d} \frac{\mu^2_{\alpha,-1} - \mu^2_{\alpha,+1}}{2\sigma^2_\alpha} + \ln \frac{\hat{\pi}_{+1}}{\hat{\pi}_{-1}}$$

(next page)

By exponentiating both sides, we obtain

$$\frac{P(Y=1|X=x)}{P(Y=-1|X=x)} = \exp\left(w_0 + \sum_{\alpha=1}^{d} w_\alpha x_\alpha\right)$$

$$\frac{P(Y=1|X=x)}{1-P(Y=1|X=x)} = \exp\left(w_0 + \sum_{\alpha=1}^{d} w_\alpha x_\alpha\right)$$

$$P(Y=1|X=x)\left(1 + \exp\left(w_0 + \sum_{\alpha=1}^{d} w_\alpha x_\alpha\right)\right) = \exp\left(w_0 + \sum_{\alpha=1}^{d} w_\alpha x_\alpha\right)$$

$$P(Y=1|X=x) = \frac{1}{1 + \exp\left[-\left(w_0 + \sum_{\alpha=1}^{d} w_\alpha x_\alpha\right)\right]}$$