1 High dimensional hi-jinx

1. Intra-class distance.

$$E[(X-X')^2] = E[(X^2+X'^2-2XX'] = E[X^2]+E[X'^2]-2E[X]E[X']$$
$$=\mu_1^2+\sigma^2+\mu_1^2+\sigma^2-2\mu_1^2 = 2\sigma^2$$

2. Inter-class distance.

$$E[(X-X')^2] = E[(X^2+X'^2-2XX'] = E[X^2]+E[X'^2]-2E[X]E[X']$$
$$=\mu_1^2+\sigma^2+\mu_2^2+\sigma^2-2\mu_1\mu_2$$

3. Intra-class distance, m-dimensions.

$$\begin{aligned}\mathbf{E}[\sum_{j=1}^{m}(X_j-X_j')^2] =& \sum_{j=1}^{m}E[(X_j-X_j^{'})^2]\\ =& 2m\sigma^2\end{aligned}$$

4. Inter-class distance, m-dimensions.

$$\begin{aligned}\mathbf{E}[\sum_{j=1}^{m}(X_j-X_j')^2] =& \sum_{j=1}^{m}E[(X_j-X_j^{'})^2]\\ =& \sum_{j=1}^{m}(\mu_{1j}-\mu_{2j})^2+2m\sigma^2\end{aligned}$$

5. The ratio of expected intra-class distance to inter-class distance is: $2m\sigma^2/\mu_{11}^2+\mu_{21}^2+2m\sigma^2$. As m increases towards ∞, this ratio approaches 1.

2 Double-counting the evidence

1. 5 parameters will be needed, P(Y=T), P($X_1 = T|Y = T$), $P(X_1 = T|Y = F)$, $P(X_2 = T|Y = F)$, $P(X_2 = T|Y = T)$.

2.

X_1	X_2	Y
T	T	T
T	F	T
F	T	T
F	F	F

3. Naive Bayes error rate

   (a) $P(X_1=T,X_2=T,Y=F) = P(X_1=T|Y=F)P(X_2=T|Y=F)P(Y=F) = 0.015$
   $P(X_1=T,X_2=F,Y=F) = P(X_1=T|Y=F)P(X_2=F|Y=F)P(Y=F) = 0.135$
   $P(X_1=F,X_2=T,Y=F) = P(X_1=F|Y=F)P(X_2=T|Y=F)P(Y=F) = 0.035$
   $P(X_1=F,X_2=F,Y=T) = P(X_1=F|Y=T)P(X_2=F|Y=T)P(Y=T) = 0.05.$

$$\begin{aligned}\mathit{error\ rate} =& \sum_{X_1,X_2,Y}1(Y\neq f(X_1,X_2))P(X_1,X_2,Y)\\ =& P(X_1=T,X_2=T,Y=F)+P(X_1=T,X_2=F,Y=F)+P(X_1=F,X_2=T,Y=F)+P(X_1=F\\ =& 0.235\end{aligned}$$

(b) $P(X_1 = T, Y = F) = P(X_1 = T|Y = F)P(Y = F) = 0.15$
$P(X_1 = F, Y = T) = P(X_1 = F|Y = T)P(Y = T) = 0.1$
so, error rate is 0.25.

(c) $P(X_2 = T, Y = F) = P(X_2 = T|Y = F)P(Y = F) = 0.05$
$P(X_2 = F, Y = T) = P(X_2 = F|Y = T)P(Y = T) = 0.25$
so, error rate is 0.30.

(d) The error rate is lower using X_1, X_2together, because we use more independent features to classify Y, which makes the classification more accurately.

4. KL divergence for Multinomials.

(a) No, $X_2$and X_2 are not conditionally independent given Y, since $P(X_2, X_3|Y) = P(X_2|Y)$ *and* $P(X_2|Y) = P(X_3|Y)$, so apparently, $P(X_2, X_3|Y) \neq P(X_2|Y)P(X_3|Y)$.

(b) After adding the X_2, the classification result becomes

X_1	X_2	X_3	Y
T	T	T	T
T	F	F	F
F	T	T	T
F	F	F	F

$P(X_1 = T, X_2 = T, X_3 = T, Y = F) = P(X_1 = T|Y = F)P(X_2 = T, X = T|Y = F)P(Y = F) = 0.015$
$P(X_1 = T, X_2 = F, X_3 = F, Y = T) = P(X_1 = T|Y = T)P(X_2 = F, X_2 = F|Y = T)P(Y = F) = 0.2$
$P(X_1 = F, X_2 = T, X_3 = T, Y = F) = P(X_1 = F|Y = F)P(X_2 = T, X_3 = T|Y = F)P(Y = F) = 0.035$
$P(X_1 = F, X_2 = F, X_2 = F, Y = T) = P(X_1 = F|Y = T)P(X_2 = F, X_3 = F|Y = T)P(Y = T) = 0.05$
so, error rate is 0.3.

5. This is because, $X_2$and $X_3$are actually not independent, but when using Naive Bayes to make decision, they are treated as independent.

6. No, logistic regression doesn't suffer from the same problem, since it calculate p(y|x) directly by estimating the weight.

7. Extra credit

(a) Decision rule

$$\frac{P(Y = T|X_1 = T, X_2 = F, X_3 = F)}{P(Y = F|X_1 = T, X_2 = F, X_3 = F)} \geq 1$$

$$\frac{\frac{P(X_! = T|Y = T)P(X_2 = F|Y = T)P(X_3 = F|Y = T)P(Y = T)}{P(X_1)P(X_2)P(X_3)}}{\frac{P(X_! = T|Y = F)P(X_2 = F|Y = F)P(X_3 = F|Y = T)P(Y = F)}{P(X_1)P(X_2)P(X_3)}} \geq 1$$

$$\frac{\frac{P(Y = T|X_1 = T)P(X_1 = T)P(Y = T|X_2 = F)P(X_2 = F)P(Y = T|X_3 = F)P(X_3 = F)P(Y = T)}{P(Y = T)P(Y = T)P(Y = T)}}{\frac{P(Y = T|X_1 = T)P(X_1 = T)P(Y = T|X_2 = F)P(X_2 = F)P(Y = T|X_2 = F)P(X_2 = F)P(Y = F)}{P(Y = T)P(Y = T)}} \geq 1$$

$$\frac{P(Y = T|X_1 = T)P(Y = T|X_2 = F)P(Y = T|X_3 = F)P(Y = T)}{P(Y = F|X_1 = T)P(Y = F|X_2 = F)P(Y = F|X_3 = F)P(Y = F)} \geq 1$$

$$\frac{pq^2}{(1-p)(1-q)^2} \geq 1$$

$$pq^2 \geq (1-q)^2 - p(1-q)^2$$

$$p \geq \frac{(1-q)^2}{q^2 + (1-q)^2}$$

(b) Real decision rule

$$\begin{aligned}\frac{P(Y=T|X_1=T,X_2=F,X_3=F)}{P(Y=F|X_1=T,X_2=F,X_3=F)} &> 1\\ \frac{P(Y=T|X_1=T)P(Y=T|X_2=F)P(Y=T)}{P(Y=F|X_1=T)P(Y=F|X_2=F))P(Y=F)} &> 1\\ \frac{pq}{(1-p)(1-q)} &> 1\\ pq &> (1-q)-p(1-q)\\ p &> 1-q\end{aligned}$$

3 Conditional independence in probability models

1. We can write $p(x_i) = \sum_{j=1}^k p(x_i|z_i = j)p(z_i = j) = \sum_{j=1}^k f_j(x_i)\pi_j$ because the probablity of x take the value of x_{i} equals the sum of probablity of x=x_{i} in every possible distribution, which is marginalization.

2. The formula for $p(x_1, \ldots, x_n)$ is $\prod_{i=1}^n \sum_{j=1}^k f_j(x_i)\pi_j$ by the derivation below (simply by applying the chain rule).

$$\begin{aligned}p(x_1,\ldots,x_n) &= \prod_{i=1}^n p(x_i)\\ &= \prod_{i=1}^n \sum_{j=1}^k f_j(x_i)\pi_j\end{aligned}$$

3. The formula for $p(z_u = v \mid x_1, \ldots, x_n)$ is $\frac{f_v(x_u)\pi_v}{\sum_{j=1}^k f_v(x_u)\pi_j}$ by the derivation below (simply by applying the bayes rule).

$$\begin{aligned}p(z_u = v \mid x_1,\ldots,x_n) &= \frac{p(x_1,\ldots,x_n \mid z_u = v)p(z_u = v)}{p(x_1,\ldots,x_n)}\\ &= \frac{f_v(x_u)\prod_{i=1}^n \sum_{j\neq u}^k f_j(x_i)\pi_v}{\prod_{i=1}^n \sum_{j=1}^k f_j(x_i)\pi_j}\\ &= \frac{f_v(x_u)\pi_v}{\sum_{j=1}^k f_v(x_u)\pi_j}\end{aligned}$$

4 Boosting.

4.1 Analyzing the training error of boosting

1. Show, $\frac{1}{m}\sum_{i=1}^m I(H(x_i) \neq y_i) \leq \frac{1}{m}\sum_{i=1}^m \exp(-f(x_i)y_i)$.
If $H(x_i) = y_i$,$I(H(x_i) \neq y_i) = 0 \leq exp(-f(x_i)y_i)$,
if $H(x_i) \neq y_i$, $-f(x_i)y_i > 0$, $I(H(x_i \neq y_i)) = 1 \leq exp(-f(x_i)y_i)$
so, $\frac{1}{m}\sum_{i=1}^m I(H(x_i) \neq y_i) \leq \frac{1}{m}\sum_{i=1}^m \exp(-f(x_i)y_i)$.

2. Show $\frac{1}{m}\sum_{i=1}^{m} exp(-f(x_i)y_i) = \prod_{t=1}^{T} Z_t$.

$$\begin{aligned}
Z_T &= \sum_{i=1}^{m} D_T(i)exp(-\alpha_T y_i h_T(x_i)) \\
&= \frac{1}{Z_{T-1}}\sum_{i=1}^{m} D_{T-1}(i)exp(-\alpha_{T-1} y_i h_{T-1}(x_i))exp(-\alpha_T y_i h_T(x_i)) \\
&= \frac{1}{Z_{T-1}}\sum_{i=1}^{m} D_{T-1}(i)\prod_{t=T-1}^{T} exp(-\alpha_t y_i h_t(x_i)) \\
&= \frac{1}{Z_{T-1}Z_{T-2}}\sum_{i=1}^{m} D_{T-2}(i)\prod_{t=T-2}^{T} exp(-\alpha_t y_i h_t(x_i)) \\
&= \frac{1}{\prod_{t=1}^{T-1} Z_t}\sum_{i=1}^{m} D_1(i)\prod_{t=1}^{T} exp(-\alpha_t y_i h_t(x_i)) \\
\implies \prod_{t=1}^{T} Z_t &= \sum_{i=1}^{m}\frac{1}{m}exp(-y_i\sum_{t=1}^{T}(\alpha_t h_t(x_i)) \\
&= \frac{1}{m}\sum_{i=1}^{m} exp(-f(x_i)y_i)
\end{aligned}$$

3. prove $Z_t = \epsilon_t exp(\alpha_t) + (1-\epsilon_t)exp(-\alpha_t)$

$$\begin{aligned}
Z_t &= \sum_{i=1}^{m} D_t(i)exp(-\alpha_t y_i h_t(x_i)) \\
&= \sum_{i\in error} D_t(i)exp(-\alpha_t y_i h_t(x_i)) + \sum_{i\in correct} D_t(i)exp(-\alpha_t y_i h_t(x_i)) \\
&= \sum_{i\in error} D_t(i)exp(\alpha_t) + \sum_{i\in correct} D_t(i)exp(-\alpha_t) \\
&= \sum_{i=1}^{m} D_t(i)I(h_t(x_i)\neq y_i)exp(\alpha_t) + \sum_{i=1}^{m} D_t(i)I(h_t(x_i)=y_i)exp(-\alpha_t) \\
&= \epsilon_t exp(\alpha_t) + (1-\epsilon_t)exp(-\alpha_t)
\end{aligned}$$

4. To minimize Z_t, differentiate it,

$$\begin{aligned}
\frac{dZ_t}{d\alpha_t} &= \epsilon_t exp(\alpha_t) - (1-\epsilon_t)exp(-\alpha_t) = 0 \\
\implies \epsilon_t exp(\alpha_t) &= (1-\epsilon_t)exp(-\alpha_t) \\
\log\epsilon_t exp(\alpha_t) &= \log(1-\epsilon_t)exp(-\alpha_t) \\
\log\epsilon_t + \alpha_t &= \log(1-\epsilon_t) - \alpha_t \\
\implies \alpha_t^* &= \frac{1}{2}\log(\frac{1-\epsilon_t}{\epsilon_t})
\end{aligned}$$

5. Replace the α_t in equation $Z_t = \epsilon_t exp(\alpha_t) + (1-\epsilon_t)exp(-\alpha_t)$ with α_t^*.

$$\begin{aligned}
Z_t &= \epsilon_t exp(\alpha_t) + (1-\epsilon_t)exp(-\alpha_t) \\
&= \epsilon_t exp(\frac{1}{2}\log(\frac{1-\epsilon_t}{\epsilon_t})) + (1-\epsilon_t)exp(-\frac{1}{2}\log(\frac{1-\epsilon_t}{\epsilon_t})) \\
&= \epsilon_t(\frac{1-\epsilon_t}{\epsilon_t})^{\frac{1}{2}} + (1-\epsilon_t)(\frac{1-\epsilon_t}{\epsilon_t})^{-\frac{1}{2}} \\
&= 2\sqrt{\epsilon_t(1-\epsilon_t)}
\end{aligned}$$

6. Prove $Z_t \leq \exp(-2\gamma_t^2)$.

$$\begin{aligned}
Z_t &= 2\sqrt{\epsilon_t(1-\epsilon_t)} \\
&= \exp(\log 2\sqrt{\epsilon_t(1-\epsilon_t)}) \\
&= \exp(\log \sqrt{4(\frac{1}{2}-\gamma_t)(\frac{1}{2}+\gamma_T)}) \\
&= \exp(\frac{1}{2}\log(1-4\gamma_t^2)) \\
&\leq \exp(\frac{1}{2} * -4\gamma_t^2) \\
&= \exp(-2\gamma_t^2) \qquad (0 < \epsilon_t = \frac{1}{2} - \gamma_t \leq 1)
\end{aligned}$$

7. Suppose not, suppose there doesn't exist such classifier, then there must be a classifer $h_t^{'}$ that its training error on the weighted dataset $\epsilon_t^{'} \geq 0.5$, then we just set the $h_t = -h_t^{'}$, the training error of which will achieve$\epsilon_t \leq 0.5$, generating contradiction, so there always exists such weak classifier.

8. Prove the training error may stuck.

$$\begin{aligned}
\alpha_t &= \frac{1}{2}\log(\frac{1-\epsilon_t}{\epsilon_t}) \\
&= \frac{1}{2}\log(\frac{1-0.5}{0.5}) \\
&= 0 \\
D_{t+1}(i) &= \frac{D_t(i)\exp(-\alpha_i y_i h_i(x_i))}{Z_t} \\
&= \frac{D_t(i)\exp(0)}{2\sqrt{\epsilon_t(1-\epsilon_t)}} \\
&= D_t(i)
\end{aligned}$$

As we can see, since $D_t(i)$ stucks, then $\epsilon_t = \sum_{i=1}^{m} D_t(i)I(h_t(x_i) \neq y_i)$ also stucks.

4.2 Adaboost on a toy dataset.

1.

ϵ_t	α_t	Z_t	$D_t(1)$	$D_t(2)$	$D_t(3)$	$D_t(3)$
0.25	0.55	0.87	0.25	0.25	0.25	0.25
0.17	0.80	0.75	0.17	0.17	0.17	0.50
0.10	1.10	0.60	0.10	0.10	0.50	0.30
0.06	1.42	0.46	0.06	0.50	0.28	0.17

2. The training error is 0.
$h = sign(\sum_t \alpha_t h_t) = sign(-1.036, 1.672, 2.26, -2.77) = -1, 1, 1, -1$

3. The dataset above is not linearly separable. Decision stump try to separte the dataset linearly which is not possible, so there must be training errors, but on the other hand, the Adaboost is trying to reweighting the dataset to make it linear separable, so that we can achieve 0 training error.