Problem 1

(a)

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

The number of parameters to estimate is $2 \times 2 \times 2 + 2 = 10$

(b) For the input has 100 features, and each feature and the label are binary

$$Combination(X) = 2^{100}$$
$$Combination(Y) = 2$$
$$Combination(X|Y) = 2 * 2^{100} = 2^{101}$$
$$\#parameters = 2^{101} + 2$$

(c) When the input has a large dimension, it is not feasible to estimate the parameters without naïve Bayesian assumption because the number of parameters to estimate grows exponentially with respect to the dimension.

For example, with 100 dimensional data and naïve Bayesian assumption, the number of parameters to estimate is

Problem 2

(a) Bayes is linear

Since we used the naïve assumption to find the maximum likelihood

The logarithmic likelihood would be

$$h(d) = \arg\max\left(\log\left(\Pr(Y = y)\right) + \sum \log(Pr(W = w_i|Y = y))\right)$$

$$h(d) = \log(\Pr(Y = y)) + \sum_i \sum_{all\ w_i} x_i \log\big(Pr(W = w_i|Y = y)\big)$$

Therefore we have

$$b = \log(\Pr(Y = y))$$

Let $v = \{v_1, v_2, \dots, v_m\}$

$$v_i = \sum_{all\ w_i} \log\big(Pr(W = w_i|Y = y)\big)$$

(b) Continues X

i. Write down MLE

$$\hat{\theta}_{\alpha c} = \frac{\sum_{i=1}^{n} I(y_i = c)[x_i]_\alpha + l}{\sum_{i=1}^{n} I(y_i = c) \sum_{\beta=1}^{d} [x_i]_\beta + ld}$$

ii. Gaussian MLE

$$\hat{\mu}_{\alpha c} \leftarrow \frac{1}{n_c} \sum\nolimits_{i=1}^{n} I(y_i = c)[x_i]_\alpha$$

$$\hat{\sigma}_\alpha^2 \leftarrow \frac{1}{n_c} \sum\nolimits_{i=1}^{n} I(y_i = c)([x_i]_\alpha - \mu_{\alpha c})^2$$

Where $n_c = \sum_{i=1}^{n} I(y_i = c)$

(c) Weight

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

$$= \frac{1}{1 + \exp\left(-\log\left(\frac{P(X|y = 1)P(y = 1)}{P(X|y = -1)P(y = -1)}\right)\right)} = \sigma\left(\sum \log\left(\frac{P(X|y = 1)}{P(X|y = -1)}\right) + \log\left(\frac{P(y = 1)}{P(y = -1)}\right)\right)$$

$$\sum \log\left(\frac{P(X|y = 1)}{P(X|y = -1)}\right) + \log\left(\frac{P(y = 1)}{P(y = -1)}\right) = \sum w_i X_i + w_o$$

Because we assume that it is a Gaussian Naïve Bayes and y is multinomial distributed

$$\sum \log\left(\frac{P(X|y=1)}{P(X|y=-1)}\right) = \sum \frac{\frac{1}{\sqrt{2\pi\sigma_i^2}}\exp(-(X_i-\mu_{i1})^2/2\sigma^2)}{\frac{1}{\sqrt{2\pi\sigma_i^2}}\exp(-(X_i-\mu_{i-1})^2/2\sigma^2)} = \sum\left(\frac{(X_i-\mu_{i-1})^2-(X_i-\mu_{i1})^2}{2\sigma_i^2}\right)$$

$$= \sum\left(\frac{\left(X_i^2-2X_i\mu_{i-1}+\mu_{i-1}^2\right)-\left(X_i^2-2X_i\mu_{i1}+\mu_{i1}^2\right)}{2\sigma_i^2}\right) = \sum\left(\frac{\mu_{i1}-\mu_{i-1}}{\sigma_i^2}X_i+\frac{\mu_{i-1}^2-\mu_{i1}^2}{2\sigma_i^2}\right)$$

We then have

$$w_i = \frac{\mu_{i1}-\mu_{i-1}}{\sigma_i^2}$$

Problem 3

(a) Because $x \in R^3 \rightarrow x = (x_1, x_2, x_3)$

$$\phi(x_i) = \left[1, \sqrt{2}(x_i)_1, \sqrt{2}\,(x_i)_2, \sqrt{2}\,(x_i)_3, (x_i)_1^2, (x_i)_2^2, (x_i)_3^2, \sqrt{2}\,(x_i)_1(x_i)_2, \sqrt{2}\,(x_i)_1(x_i)_3, \sqrt{2}\,(x_i)_2(x_i)_3\right]$$

$$\langle \phi(x_i), \phi(x_j) \rangle = 1 + 2(x_i)_1\left(x_j\right)_1 + 2(x_i)_2\left(x_j\right)_2 + 2(x_i)_3\left(x_j\right)_3$$

$$+(x_i)_1^2\left(x_j\right)_1^2 + (x_i)_2^2\left(x_j\right)_2^2 + (x_i)_3^2(x_3)_3^2$$

$$+2(x_i)_1\left(x_j\right)_1(x_i)_2\left(x_j\right)_2 + 2(x_i)_1\left(x_j\right)_1(x_i)_3\left(x_j\right)_3 + 2(x_i)_2\left(x_j\right)_2(x_i)_3\left(x_j\right)_3$$

There are ten elements in $\phi(x_i)$ therefore the D is ten.

If we do explicit feature mapping, we would first calculate the mapped values and then use the mapped value to calculate the inner product. This feature transformation and inner product process may require a lot of computation. Because we have the input x, we can directly evaluate the inner product – skipping the explicit mapping and hence saving the computation.

(b) Positive semidefinite matrix

Let v_1 be any 2×1 column vector

$$v_1^T A_1 v_1 = \begin{pmatrix} v_1^1 & v_2^1 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} v_1^1 \\ v_2^1 \end{pmatrix} = \begin{pmatrix} v_1^1 + v_2^1 & v_1^1 + v_2^1 \end{pmatrix}\begin{pmatrix} v_1^1 \\ v_2^1 \end{pmatrix}$$

$$v_1^T A_1 v_1 = v_1^1(v_1^1 + v_2^1) + v_2^1(v_1^1 + v_2^1) = (v_1^1 + v_2^1)(v_1^1 + v_2^1) = (v_1^1 + v_2^1)^2 \geq 0$$

Therefore matrix A_1 is positive semidefinite.

Let v_2 be any 3×1 column vector

$$v_2^T A_2 v_2 = \begin{pmatrix} v_1^2 & v_2^2 & v_3^2 \end{pmatrix}\begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 0 \\ 1 & 0 & 2 \end{pmatrix}\begin{pmatrix} v_1^2 \\ v_2^2 \\ v_2^2 \end{pmatrix} = \begin{pmatrix} 2v_1^2 + v_2^2 + v_3^2 & v_1^2 + 2v_2^2 & v_1^2 + 2v_3^2 \end{pmatrix}\begin{pmatrix} v_1^2 \\ v_2^2 \\ v_2^2 \end{pmatrix}$$

$$v_2^T A_2 v_2 = v_1^2(2v_1^2 + v_2^2 + v_3^2) + v_2^2(v_1^2 + 2v_2^2) + v_3^2(v_1^2 + 2v_3^2)$$

$$v_2^T A_2 v_2 = 2(v_1^2)^2 + 2(v_2^2)^2 + 2(v_3^2)^2 + 2v_1^2 v_2^2 + 2v_1^2 v_3^2 = (v_1^2 + v_2^2)^2 + (v_1^2 + v_3^2)^2 + (v_2^2)^2 + (v_3^2)^2$$

$$v_2^T A_2 v_2 > 0$$

Therefore matrix A_2 is strictly positive definite with non-zero v_2.

Let v_3 be any 3×1 column vector

$$v_3^T A_3 v_3 = \begin{pmatrix} v_1^3 & v_2^3 & v_3^3 \end{pmatrix}\begin{pmatrix} 2 & 1 & -1 \\ 1 & 1 & 1 \\ -1 & 1 & 2 \end{pmatrix}\begin{pmatrix} v_1^3 \\ v_2^3 \\ v_2^3 \end{pmatrix}$$

$$= \begin{pmatrix} 2v_1^3 + v_2^3 - v_3^3 & v_1^3 + v_2^3 + v_3^3 & -v_1^3 + v_2^3 + 2v_3^3 \end{pmatrix}\begin{pmatrix} v_1^3 \\ v_2^3 \\ v_2^3 \end{pmatrix}$$

$$v_3^T A_3 v_3 = v_1^3(2v_1^3 + v_2^3 - v_3^3) + v_2^3(v_1^3 + v_2^3 + v_3^3) + v_3^3(-v_1^3 + v_2^3 + 2v_3^3)$$

$$\begin{aligned} v_3^T A_3 v_3 &= 2(v_1^3)^2 + (v_2^3)^2 + 2(v_3^2)^2 + 2v_1^3 v_2^3 - 2v_1^2 v_3^2 + 2v_2^3 v_3^3 \\ &= (v_1^3 + v_2^3)^2 + (v_1^3 - v_3^3)^2 + (v_2^3 + v_3^3)^2 - (v_2^3)^2 \end{aligned}$$

$v_3^T A_3 v_3$ is not guaranteed to be non-negative or non-positive

Therefore matrix A_3 is neither positive semidefinite or strictly positive definite.

(c) Let A and B be $n \times n$ psd matrices, then we have

$$v^T A v \geq 0$$

$$v^T B v \geq 0$$

Since $A + B = \begin{pmatrix} a_{11} + b_{11} & \cdots & a_{1n} + b_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} + b_{n1} & \cdots & a_{nn} + b_{nn} \end{pmatrix}$

$$v^T(A + B)v = v^T A v + v^T B v \geq 0$$

Therefore $A + B$ is also positive semidefinite.

A matrix is positive semidefinite if and only if all of its eigenvalues are non-negative. Because both A and B are positive semidefinite, their eigenvalues are all non-negative.

Let $\{\lambda_1, \lambda_2, \dots, \lambda_p\}$ be the eigenvalues of matrix A and $\{\mu_1, \mu_2, \dots, \mu_q\}$ be the eigenvalues of matrix B. By the definition of positive semidefinite

$$\lambda_p \geq 0, \forall p$$

$$\mu_q \geq 0, \forall q$$

There is pq corresponding eigenvectors for the Kronecker product $A \otimes B = \begin{pmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{n1}B & \cdots & a_{nn}B \end{pmatrix}$ and they are $\lambda_p \mu_q$ for all values of p and q. Because we have

$$\lambda_p \geq 0, \forall p$$

$$\mu_q \geq 0, \forall q$$

$$\rightarrow \lambda_p \mu_q \geq 0, \forall p, q$$

Therefore, the Kronecker product of two positive semidefinite matrices is also positive semidefinite.

(d) RBF kernel $K(x, x') = \exp(-\frac{\left||x - x'|\right|^2}{2\sigma^2})$

Suppose that $x \in R^2$ such that $x = (x_1, x_2)$. Assume that $2\sigma^2 = 1$

$$K(x, x') = \exp\left(-\left||x - x'|\right|^2\right) = \exp\left(-\left||x|\right|^2\right) \exp\left(-\left||y|\right|^2\right) \exp(2x^T y)$$

Using the Taylor series

$$K(x, x') = \exp\left(-\left|\left|x\right|\right|^2\right) \exp\left(-\left|\left|y\right|\right|^2\right) \sum_{n=0}^{\infty} \frac{(2x^T y)^n}{n!}$$

That is, RBF kernel corresponds to the inner product in an infinite dimensional space

Problem 4

(a) The linear combination is

$$w = \sum_{i}^{n} \alpha_i y_i x_i$$

Where α_i is the number of times the training input x_i is misclassified because each time a misclassification will result in the weight update.

(b) Because for the perceptron learning we have $\hat{y} = sign(w^T x)$

$$\hat{y} = sign\left(\sum_{i}^{n} \alpha_i y_i x_i\right)^T x$$

$$h(x) = sign\left(\sum_{i}^{n} \alpha_i y_i (x_i, x)\right) = sign\left(\sum_{i}^{n} \alpha_i y_i K(x_i, x)\right)$$

(c) Suppose we have n training inputs

1. Initialize the α to an all-zeros vector of size n.
2. Loop until the maximum number of iteration or other stopping criteria reached
   For each piece training data (x_j, y_j), calculate

$$\hat{y} = sign\left(\sum_{i}^{n} \alpha_i y_i K(x_i, x_j)\right)$$

   If $\hat{y} \neq y_j$, then update the misclassification vector

$$\alpha_j = \alpha_j + 1$$