CHUNYUAN LI
JIARUI XING

CSE 517A
Machine Learning
THW2

Problems

1. (30 points) Naïve Bayes

(a) We have 2^2 different $\vec{x}$s under 2 different categories ys, so in total we need $2^2 * 2 = 8$ parameters for P(X|Y).

(b) This time we need $2^d * 2 = 2^{d+1} = 2^{101}$ parameters.

(c) If we make the naïve assumption, for each dimension X_i we need only to estimate $P(X_i|Y = 0)$ and $P(X_i|Y = 1)$, therefore in total we need $2d$ parameters, which is much smaller than that of not using naïve assumption (2^{d+1}). Generally speaking, the naïve assumption can be seen as an approximation method which makes our problem easier to solve. So when the dimension is high, we'd better use it or the problem is almost intractable for having too much (exponential # of) parameters to estimate. However, in lower dimension where we are able to estimate all the parameters, we can get more precise result by not using it.

2. (30 points) Naïve Bayes, Part II

(a) Use $\vec{x}$ as feature vector of document, we have:

$$h(d) = \underset{y}{argmax}\, P(Y=y)\prod_{i=1}^{l} P(W_i = w_i|Y=y)$$

$$= \underset{y}{argmax}\, P(Y=y)\prod_{i=1}^{d} P(X_i = x_i|Y=y)$$

$$= \underset{y}{argmax}\, \mathrm{P(Y=y)}\prod_{i=1}^{d} P(a_i \in d|Y=y)^{x_i}$$

$$= \underset{y}{argmax}\ log\left(\mathrm{P(Y=y)}\prod_{i=1}^{d} P(a_i \in d|Y=y)^{\ x_i}\right)$$

$$= \underset{y}{argmax}\ logP(Y=y) + \sum_{i=1}^{d} log[P(a_i \in d|Y=y)]x_i$$

$$= sign\left[logP(Y=1) + \sum_{i=1}^{d} log[P(a_i \in d|Y=1)]x_i - logP(Y=-1) - \sum_{i=1}^{d} log[P(a_i \in d|Y=-1)]x_i\right]$$

$$= \mathrm{sign}\left[\sum_{\mathrm{i=1}}^{\mathrm{d}} \log\frac{P(a_i \in d|Y=1)}{P(a_i \in d|Y=-1)}x_i + log\frac{P(Y=1)}{P(Y=-1)}\right]$$

Therefore,

$$\vec{v} = \left(\log\frac{P(a_1 \in d|Y=1)}{P(a_1 \in d|Y=-1)}, \log\frac{P(a_2 \in d|Y=1)}{P(a_2 \in d|Y=-1)}, \dots, \log\frac{P(a_d \in d|Y=1)}{P(a_d \in d|Y=-1)}\right)$$

$$\mathrm{b} = \log\frac{\mathrm{P(Y=1)}}{\mathrm{P(Y=-1)}}$$

(b) (i) Assuming we have in total n samples, X_i has K_i different categories and Y has K_y different classes,

$$Pr(X_i|Y) = \prod_{j=1}^{K_i} Pr(X_i = j \mid y)^{I(X_i=j)}$$

$$Pr(Y=c) = \frac{\sum_{i=1}^{K} I(y_i = c)}{K}$$

(ii) The MLE estimators for Gaussian distribution are:

$$\mu_i^* = \bar{X}_\iota = \frac{1}{N}\sum_{j=1}^{N} X_{ij}$$

$$\sigma_i^* = \frac{1}{N}$$

(iii) We have 2 classes, and for each we need to estimate d means, $d$variances and 1 prior $p(Y)$, so there should be $2 * (d + d + 1) = 4d + 2$ parameters.

(c)

$$\because \mathrm{P(X|Y=1)} = \prod_{\mathrm{i=1}}^{\mathrm{d}} \mathrm{P(X_i|Y=1)}$$

$$= \prod_{\mathrm{i=1}}^{\mathrm{d}} \frac{1}{\sqrt{2\pi}\sigma_i} exp\left(-\frac{\left(X_i - \mu_{1,i}\right)^2}{2\sigma_{\mathrm{i}}{}^2}\right)$$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^{\mathrm{d}} \exp\left(\sum_{\mathrm{i}=1}^{\mathrm{d}} \frac{1}{2{\sigma_{\mathrm{i}}}^2}(X_{\mathrm{i}} - \mu_1)^2\right) \prod_{i=1}^{d} \frac{1}{\sigma_i}$$

$$\therefore \mathrm{P}(\mathrm{Y} = +1|\mathrm{X}) = \frac{\mathrm{P}(\mathrm{X}|\mathrm{Y} = +1)\mathrm{P}(\mathrm{Y} = +1)}{\mathrm{P}(\mathrm{X})}$$

$$= \frac{\mathrm{P}(\mathrm{Y} = +1)\left(\frac{1}{\sqrt{2\pi}}\right)^{\mathrm{d}} \exp\left(\sum_{\mathrm{i}=1}^{\mathrm{d}} \frac{1}{2{\sigma_{\mathrm{i}}}^2}(X_{\mathrm{i}} - \mu_1)^2\right) \prod_{i=1}^{d} \frac{1}{\sigma_i}}{\mathrm{P}(\mathrm{Y} = +1)\left(\frac{1}{\sqrt{2\pi}}\right)^{\mathrm{d}} \exp\left(\sum_{\mathrm{i}=1}^{\mathrm{d}} \frac{1}{2{\sigma_{\mathrm{i}}}^2}(X_{\mathrm{i}} - \mu_1)^2\right) \prod_{i=1}^{d} \frac{1}{\sigma_i} + \mathrm{P}(\mathrm{Y} = -1)\left(\frac{1}{\sqrt{2\pi}}\right)^{\mathrm{d}} \exp\left(\sum_{\mathrm{i}=1}^{\mathrm{d}} \frac{1}{2{\sigma_{\mathrm{i}}}^2}\left(X_{\mathrm{i}} - \mu_{(-1)}\right)^2\right) \prod_{i=1}^{d} \frac{1}{\sigma_i}}$$

$$= \frac{1}{1 + \frac{\mathrm{P}(\mathrm{Y} = -1)}{\mathrm{P}(\mathrm{Y} = +1)} \exp\left(\sum_{\mathrm{i}=1}^{\mathrm{d}} [\frac{1}{{\sigma_{\mathrm{i}}}^2}(\mu_1 - \mu_{(-1)})X_{\mathrm{i}} + \frac{1}{2\sigma_i^2}(\mu_{(-1)}^2 - \mu_1^2)\right)}$$

$$= \frac{1}{1 + \exp\left(\log\mathrm{P}(\mathrm{Y} = -1) - \log\mathrm{P}(\mathrm{y} = +1) + \sum_{1}^{\mathrm{d}} \frac{1}{2\sigma_i^2}(\mu_{(-1)}^2 - \mu_1^2) + \sum_{\mathrm{i}=1}^{\mathrm{d}} \frac{1}{{\sigma_{\mathrm{i}}}^2}(\mu_{(-1)} - \mu_1)X_{\mathrm{i}}\right)}$$

As a result,

$$\mathrm{w_i} = \frac{1}{{\sigma_{\mathrm{i}}}^2}(\mu_1 - \mu_{(-1)}), i = 1,2, \dots, d$$

$$w_0 = logP(Y = +1) - logP(y = -1) + \sum_{1}^{\mathrm{d}} \frac{1}{2\sigma_i^2}(\mu_{(-1)}^2 - \mu_1^2) = log\frac{P(Y = +1)}{P(Y = -1)} + \sum_{1}^{\mathrm{d}} \frac{1}{2\sigma_i^2}(\mu_{(-1)}^2 - \mu_1^2)$$

3. (40 points) Valid Kernels, Kernel Construction

(a) Here $D = 10$. Define kernel function K(·,·):

$$K(\boldsymbol{x_i}, \boldsymbol{x_j}) = \left(1 + \boldsymbol{x_i^T x_j}\right)^2$$

$$= \left(1 + \sum_{d=1}^{3} (x_i)_d \left(x_j\right)_d\right)^2$$

$$= 1 + 2\sum_{\mathrm{d}=1}^{3} (\mathrm{x_i})_\mathrm{d} \left(\mathrm{x_j}\right)_\mathrm{d} + 2\sum_{\mathrm{d}=1}^{3} (x_i)_d^2 \left(x_j\right)_d^2 + 2(x_i)_1 (x_i)_2 \left(x_j\right)_1 \left(x_j\right)_2 + 2(x_i)_2 (x_i)_3 \left(x_j\right)_2 \left(x_j\right)_3$$

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

$$=< \Phi(x_i), \Phi\left(x_j\right) >$$

We can see that transforming $\boldsymbol{x}$ to $\Phi(\boldsymbol{x})$ needs 12 productions, and compute $< \Phi(x_i), \Phi\left(x_j\right) >$ need additional 10 productions and 9 summations, thus in total we need 22 productions and 9 summations. However, compute $K(\boldsymbol{x_i}, \boldsymbol{x_j}) = \left(1 + \boldsymbol{x_i^T x_j}\right)^2$ need only 4 productions and 1 summation and can get the same result, which is highly efficient.

(b) We can judge from the eigenvalues of matrices.

(i) For $\mathrm{A}_1 = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$, its eigen vectors and eigenvalues are:

$$v_1 = (1,1)^T, \lambda_1 = 2 \geq 0$$
$$v_2 = (-1,1)^T, \lambda_2 = 0 \geq 0$$

Therefore, A_1 is positive semidefinite.

(ii) For $\mathrm{A}_2 = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 0 \\ 1 & 0 & 2 \end{bmatrix}$, its eigenvectors and eigenvalues are:

$$\mathrm{v}_1 = \left(\sqrt{2}, 1, 1\right)^{\mathrm{T}}, \lambda_1 = 2 + \sqrt{2} > 0$$

$$v_2 = (0, -1, -1)^T, \lambda_2 = 2 > 0$$

$$\mathrm{v}_3 = \left(-\sqrt{2}, -1, 1\right)^{\mathrm{T}}, \lambda_3 = 2 - \sqrt{2} > 0$$

Therefore, A_2 is positive definite.

(iii) For $\mathrm{A}_3 = \begin{bmatrix} 2 & 1 & -1 \\ 1 & 1 & 1 \\ -1 & 1 & 2 \end{bmatrix}$, its eigenvectors and eigenvalues are:

$$\mathrm{v}_1 = (-1,0,1)^{\mathrm{T}}, \lambda_1 = 3 > 0$$

$$\mathrm{v}_2 = \left(1, \frac{-2+\sqrt{2}}{-1+\sqrt{2}}, 1\right)^{\mathrm{T}}, \lambda_2 = 1 + \sqrt{2} > 0$$

$$\mathrm{v}_3 = \left(1, -\frac{2+\sqrt{2}}{1+\sqrt{2}}, 1\right)^{\mathrm{T}}, \lambda_3 = 1 - \sqrt{2} < 0$$

Therefore, A_1 is neither positive definite or semidefinite.

(c) (i) Assuming A and B are both $\mathrm{n} \times \mathrm{n}$ matrices, since A and B are both positive semi-definite matrices, we have:

$$\forall \lambda \in \mathrm{R}^{\mathrm{n}}, \lambda^{\mathrm{T}} \mathrm{A} \lambda \geq 0, \lambda^{\mathrm{T}} \mathrm{B} \lambda \geq 0.$$

Therefore,

$$\forall \lambda \in R^n, \lambda^{\mathrm{T}}(\mathrm{A}+\mathrm{B})\lambda = \lambda^T A\lambda + \lambda^T B\lambda \geq 0$$

Thus (A+B) is positive semi-definite.

(ii) Using property of Kronecker product [1]:

Suppose that A and B are square matrices of size n and m respectively. Let $\lambda_1, \ldots, \lambda_{\mathrm{n}}$ be the eigenvalues of A and $\mu_1, \ldots, \mu_{\mathrm{m}}$ be those of B (listed according to multiplicity). Then the eigenvalues of A $\otimes$ B are

$$\lambda_{\mathrm{i}}\mu_{\mathrm{j}}, \qquad \mathrm{i} = 1, \ldots, \mathrm{n},\ \mathrm{j} = 1, \ldots, \mathrm{m}.$$

Since A and B are both positive semi-definite, we know:

$$\lambda_1, \ldots, \lambda_n \geq 0, \qquad \mu_1, \ldots, \mu_m \geq 0$$

Therefore, $\lambda_i\mu_j \geq 0, i = 1, \ldots, n, j = 1, \ldots, m,$ i.e. all eigenvalues of A $\otimes$ B nonnegative. As a result A $\otimes$ B is positive semi-definite.

(d) By Taylor expansion:

$$\mathrm{K}(\mathrm{x_i}, \mathrm{x_j}) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$$

$$= \sum_{n=0}^{\infty} \frac{\left(x_i^T x_j\right)^n}{j!\,(2\sigma)^n} exp\left(-\frac{1}{2\sigma^2}\|x_i\|^2\right) exp\left(-\frac{1}{2\sigma^2}\|x_j\|^2\right)$$

$$= exp\left(-\frac{1}{2\sigma^2}\|x_i\|^2\right) exp\left(-\frac{1}{2\sigma^2}\|x_j\|^2\right) \sum_{n=0}^{\infty} \frac{\left(x_i^T x_j\right)^n}{n!\,(2\sigma)^n}$$

$$= \Phi(x_i)^T \Phi(x_j)$$

Where

$$\Phi(x) = exp\left(-\frac{1}{2\sigma^2}\|x\|^2\right)\left[1, \frac{x}{1!\,(2\sigma)^1}, \ldots, \frac{x^n}{n!\,(2\sigma)^n}, \ldots\right]$$

Which is a vector in infinite dimensional space.

References:

[1] Kronecker product - Wikipedia

4. (30 points) Kernelize the Perceptron Algorithm

(a) Note where α_i is the number of times $\boldsymbol{x}_i$ was misclassified, then:

$$\boldsymbol{w} = \sum_{i}^{n} \alpha_i y_i \boldsymbol{x}_i$$

(b) From (a):

$$h(x) = sign(w^T x)$$

$$= sign\left[\left(\sum_{i}^{n} \alpha_i y_i \boldsymbol{x_i}\right)^T \boldsymbol{x}\right]$$

$$= sign\left[\sum_{i=1}^{n} \alpha_i y_i \left(\boldsymbol{x}_i^T \boldsymbol{x}\right)\right]$$

If we use feature transformation $\Phi(\cdot)$, and kernel function $K(\boldsymbol{x}_i, \boldsymbol{x_j}) =< \Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x}_j) >$, then:

$$h(\boldsymbol{x}) = sign\left[\sum_{i=1}^{n} \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x})\right]$$

(c) The kernelized version of perceptron is:

(0) Initialize $\mathrm{w} = 0$
REPEAT until convergence:
(1) Pick $(\boldsymbol{x_j}, y_j)$ randomly from D
(2) If $h(\boldsymbol{x}_j) = sign\left[\sum_{i=1}^{n} \alpha_i y_i K(\boldsymbol{x_i}, \boldsymbol{x_j})\right] \neq y_j$, update $\mathrm{a_j}$: $\mathrm{a_j} \leftarrow \mathrm{a_j} + 1$