Week 7:

1) Naive Bayes classifiers:

~~lets~~ In Naive Bayes classification, we assume features are independent given class.

$$\cancel{p(x/y=c) \propto p(x, y=c)}$$

$$p(y=c/x) \propto p(x, y=c) \quad (\because \text{Generative model})$$

$$\propto p(x/y=c)\, p(y=c)$$

$$\propto \prod_{i=1}^{D} p(x_i/y=c) \cdot p(y=c)$$

MLE for Naive Bayes:

Lets write likelihood for single data case,

$$p(x_i, y_i) = \prod_{j=1}^{D} p(x_{ij}/y_i) \cdot p(y_i)$$

$$= \prod_{j=1}^{D} \prod_{c=1}^{C} p(x_{ij}/y_i=c)^{1\{y_i=c\}} \cdot \prod_{c=1}^{C} p(y_i=c)^{1\{y_i=c\}}$$

log likelihood

$$\log p(\theta) = \sum_{i=1}^{N} \log\left(\sum_{j=1}^{D} \sum_{c=1}^{C} p(x_{ij}/y_i=c)^{1\{y_i=c\}} \cdot \prod_{c=1}^{C} p(y_i=c)^{1\{y_i=c\}} \right)$$

$$= \sum_{c=1}^{C} N_c \log \pi_c + \sum_{j=1}^{D} \sum_{c=1}^{C} \sum_{i;\, y_i=c} p(x_{ij}/y_i=c) \rightarrow (1)$$

[Here we assumed $P(y)$ follows multi~~nomial~~noulli distribution with $p(y_i=c) = \pi_c$]

and each feature in 'x_j' can take different kind of distribution. So, for simplicity lets assume, it follows Bernoulli i.e $p(x_{ij}/y_i = c) \sim Ber(\theta_{jc})$

~~$\frac{\partial(\log p(\theta))}{\partial \pi_c} = \frac{\partial}{\partial \pi_c}\left(\sum_{c=1}^{C} N_c \log \pi_c\right)$~~

Then $\hat{\pi}_x, \hat{\theta}$ can be derived by

$$\underset{\hat{\pi}, \hat{\theta}}{\operatorname{argmax}} \log p(\theta) = \underset{\hat{\pi}, \hat{\theta}}{\operatorname{argmax}} \text{ — Expr. } (1)$$

when we are maximizing w.r.t $\hat{\pi}$, $\hat{\theta}$ will be const

so $\hat{\pi}_{MLE} = \underset{\pi}{\operatorname{argmax}} \sum_{c=1}^{C} N_c \log \pi_c$.

Using MLE derivation for multinomial, this

become $\hat{\pi}_{c\,MLE} = \frac{N_c}{N} = \frac{N_c}{\sum_{c=1} N_c}$

and in same way $\hat{\theta}_{jc} = \frac{N_{jc}}{N_c}$

For Baysian Model averaging, we take prior also into account. if we assume prior for π as Dirichlet(α) and for each θ_{jc} as Beta(β_0, β_1) then prior

$$~~p(\theta/D)~~ \; p(\theta) = p(\pi) \prod_{j=1}^{D} \prod_{c=1}^{C} p(\theta_{jc}) \;(\because \text{factored prior})$$

$$p(\pi) \sim Dir(\alpha), \quad p(\theta_{jc}) \sim Beta(\beta_0, \beta_1)$$

then posterior

$$p(\theta/D) = p(\pi/D) \prod_{j=1}^{D} \prod_{c=1}^{C} p(\theta_{jc}/D)$$

$p(\pi/D)$ is posterior for Dirichlet-Multinomial which is $p(\pi/\mathcal{D}) = Dir(N_1+\alpha_1, \ldots N_c+\alpha_c)$

and $p(\theta_{jc}/\mathcal{D})$ is posterior of Beta-bernoulli model

$$p(\theta_{jc}/\mathcal{D}) = Beta(N_{jc}+\beta_0, (N_{c}-N_{jc})+\beta_1)$$

the Maximum Aposteriori (MAP) estimates for them are the ~~mean of~~ the expectation of them (mean)

So $$\hat{\pi}_{c_{MAP}} = \frac{N_c+\alpha_c}{N+\alpha} \quad (\alpha = \sum_i \alpha_i)$$

and $$\hat{\theta}_{jc_{MAP}} = \frac{N_{jc}+\beta_0}{N_c+\beta_0+\beta_1}$$

- Here, MAP can be seen as single point estimates for posterior and MLE as single point estimate to ~~capture~~ explain likelihood

- MAP estimate is also can be seen as weighted average (convex combination) of MLE and prior eq.

1.3.1: Bayesian inference when σ^2 is unknown:

In linear regression we model $p(y/x) \sim N(w^T x, \sigma^2)$

In bayesian modeling, we set priors for w and σ^2.

$$p(w, \sigma^2) = p(w/\sigma^2)\, p(\sigma^2)$$

Lets take $p(w/\sigma^2)$ as $N(w_0, \sigma^2 V_0)$ and

$p(\sigma^2)$ as $IG(\sigma^2/a_0, b_0)$.

Joint Prior:

$$p(w, \sigma^2) = N(w/w_0, \sigma^2 V_0)\, IG(\sigma^2/a_0, b_0)$$

$$= \frac{1}{(2\pi)^{D/2} |\sigma^2 V_0|^{1/2}} \cdot e^{-\frac{1}{2}(w-w_0)^T (\sigma^2 V_0)^{-1} (w-w_0)}$$

$$\times \frac{b_0^{a_0}}{\Gamma(a_0)} \cdot \sigma^{2^{-(a_0+1)}} \cdot e^{-b_0/\sigma^2}$$

$$= \frac{b_0^{a_0}}{(2\pi)^{D/2} \cdot |V_0|^{1/2} \cdot \Gamma(a_0)} \cdot \sigma^{2^{-(a_0 + D/2 + 1)}} \times e^{-\frac{((w-w_0)^T V_0^{-1}(w-w_0) + 2b_0)}{2\sigma^2}}$$

likelihood:

$$P(y/x, w, \sigma^2) = N(y/Xw, \sigma^2 I_N)$$

$$= \frac{1}{(2\pi)^{D/2} \cdot |\sigma^2 I_N|^{1/2}} \cdot e^{-\frac{1}{2}(y-Xw)^T \cdot (\sigma^2 I_N)^{-1} (y-Xw)}$$

$$\propto \sigma^{2^{-D/2}} \cdot e^{-\frac{1}{2\sigma^2}(y-Xw)^T(y-Xw)}$$

Posterior for joint distribution of (w, σ^2) is

$$P(w, \sigma^2/D) = P(y/X, w, \sigma^2) \cdot P(w, \sigma^2)$$

$$\propto \sigma^{2^{-(a_0+D+1)}} \cdot e^{-\left(\frac{(y-Xw)^T(y-Xw) + (w-w_0)^T V_0^{-1}(w-w_0)}{2\sigma^2}\right)}$$

which is again in the form of Normal-Inverse Gamma distribution.

$$= NIG(w, \sigma^2 / w_N, V_N, a_N, b_N)$$

w_0 is updated to $w_N = V_N(V_0^{-1} w_0 + X^T y)$

V_0 is updated to $V_N = (V_0^{-1} + X^T X)^{-1}$

a_0 is updated to $a_N = a_0 + N/2$

b_0 is updated to $b_N = b_0 + \frac{1}{2}(w_0^T V_0^{-1} w_0 + y^T y - w_N^T V_N^{-1} w_N$

4.19) Generative classifier

$$p_\theta(y=c/x) = \frac{P_\theta(y=c)\,P_\theta(x/y=c)}{\sum_{c'} P_\theta(y=c')\,P_\theta(x/y=c')}$$

In Gaussian discriminant, we model

$$P(y) = \mathrm{Mul}(\pi) \quad P(x/y=c) = N(\mu_c, \Sigma_c)$$

$$\theta = [\pi_{1:c}, \mu_{1:c}, \Sigma_{1:c}]$$

In given problem, we have only two classes $c=2$ and $\Sigma_1 = K\Sigma_0$, $K>0$, then we can write

$$P(y=1/x) = \frac{\pi_1 \cdot |2\pi\Sigma_1|^{-1/2}\, e^{-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)}}{\sum_{i=0}^{1} \pi_i \cdot |2\pi\Sigma_i|^{-1/2}\, e^{-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)}}$$

we'll simplify numerator & denominator seperatly by substituting $\Sigma_1 = K\Sigma_0$, $K>0$

$$\text{Numerator of } p(y=1/x) = \pi_1 \cdot |2\pi\cdot(K\Sigma_0)|^{-1/2} \cdot e^{-\frac{1}{2}(x-\mu_1)(K\Sigma_0)^{-1}(x-\mu_1)}$$

$$= \frac{1}{\sqrt{K}} \cdot \pi_1 \cdot |2\pi\Sigma_0|^{-1/2} \cdot e^{-\frac{1}{2K}(x-\mu_1)\Sigma_0^{-1}(x-\mu_1)}$$

Denominator of $p(y=1/x) =$

$$= \pi_1 \cdot |2\pi(K\Sigma_0)|^{-1/2} \cdot e^{-\frac{1}{2}(x-\mu_1)(K\Sigma_0)^{-1}(x-\mu_1)} \cancel{+\pi_0 |2\pi\Sigma_0|^{-1/2}}$$

$$+\pi_0 \cdot |2\pi\Sigma_0|^{-1/2} \cdot e^{-\frac{1}{2}(x-\mu_0)\Sigma_0^{-1}(x-\mu_0)}$$

$= \cancel{\frac{1}{\sqrt{K}}(\pi_1 \cdot |2\pi\Sigma_0|^{-1}}$

$$= |2\pi\Sigma_0|^{-1/2} \cdot e^{-\frac{1}{2}x^T\Sigma_0 x} \left[\frac{\pi_1}{\sqrt{K}} \cdot e^{\frac{1}{K}(\mu_1^T\Sigma_0^{-1}x - \frac{1}{2}\mu_1^T\Sigma_0\mu_1)} + \pi_0 \cdot e^{(\mu_0^T\Sigma_0^{-1}x - \frac{1}{2}\mu_0^T\Sigma_0\mu_0)} \right]$$

After we cancel first term i.e $|2\pi\Sigma_0|^{-1/2} \cdot e^{-\frac{1}{2}x^T\Sigma_0 x}$ both in numerator & denominator we get.

$$= \frac{\left[\frac{\pi_1}{\sqrt{K}} \cdot e^{1/K(\mu_1^T\Sigma_0^{-1}x - \frac{1}{2}\mu_1^T\Sigma_0\mu_1)}\right]}{\frac{\pi_1}{\sqrt{K}} \cdot e^{1/K(\mu_1^T\Sigma_0^{-1}x - \frac{1}{2}\mu_1^T\Sigma_0\mu_1)} + \pi_0 \cdot e^{(\mu_0^T\Sigma_0^{-1}x - \frac{1}{2}\mu_0^T\Sigma_0\mu_0)}}$$

lets simplify this to by substituting

$\beta_1 = \Sigma^{-1}\mu_1$ $\qquad$ $\beta_1 = \Sigma^{-1}\mu_0$

$r_1 = -\frac{1}{2}\mu_1^T\Sigma_0\mu_1 + \log\pi_1$ $\qquad$ $r_0 = -\frac{1}{2}\mu_0^T\Sigma_0\mu_0 + \log$

$$= \frac{e^{1/K(\beta_1^T x + r_1 - \frac{\log K}{2})}}{e^{1/K(\beta_1^T x + r_1 - \frac{1}{2}\log K)} + e^{(\beta_0^T x + r_0)}}$$

$$= \frac{1}{1 + e^{\beta_0^T x + r_0 - 1/K(\beta_1^T x + r_1 + \frac{1}{2}\log K)}}$$

$$= \frac{1}{1 + e^{(\beta_0 - 1/k\,\beta_1)^T x + (r_0 - 1/k\, r_1) + \frac{\log k}{2k}}}$$

$$= \text{sigm}\left((\beta_{1/k} - \beta_0)^T x + (1/k\, r_1 - r_0) + \frac{\log k}{2k} \right).$$

So, Even $\Sigma_1 = k\Sigma_0$, the discriminative ~~surface~~ function is still sigmoid function. and boundary is still linear.