Probabilistic generative models:

Joint PDF $P(\underline{x}, y)$ or $P(\underline{x}, y = C_k)$ or $P(\underline{x}, C_k)$ gives complete summary of the uncertainity associated with $\underline{x}$ and y.

Bayes rule:

$$P(\underline{x}, C_k) = P(\underline{x} \mid C_k)\, P(C_k)$$

$$= P(C_k \mid \underline{x})\, P(\underline{x})$$

$$P(C_k \mid \underline{x}) = \frac{P(\underline{x} \mid C_k)\, P(C_k)}{P(\underline{x})}$$

$$= \frac{P(\underline{x} \mid C_k)\, P(C_k)}{\sum_{k=1}^{C} P(\underline{x} \mid C_k)\, P(C_k)}$$

$C = 2$

likelihood / class-conditional density

Prior

$$P(C_1 \mid \underline{x}) = \frac{P(\underline{x} \mid C_1)\, P(C_1)}{P(\underline{x} \mid C_1)\, P(C_1) + P(\underline{x} \mid C_2)\, P(C_2)}$$

Posterior

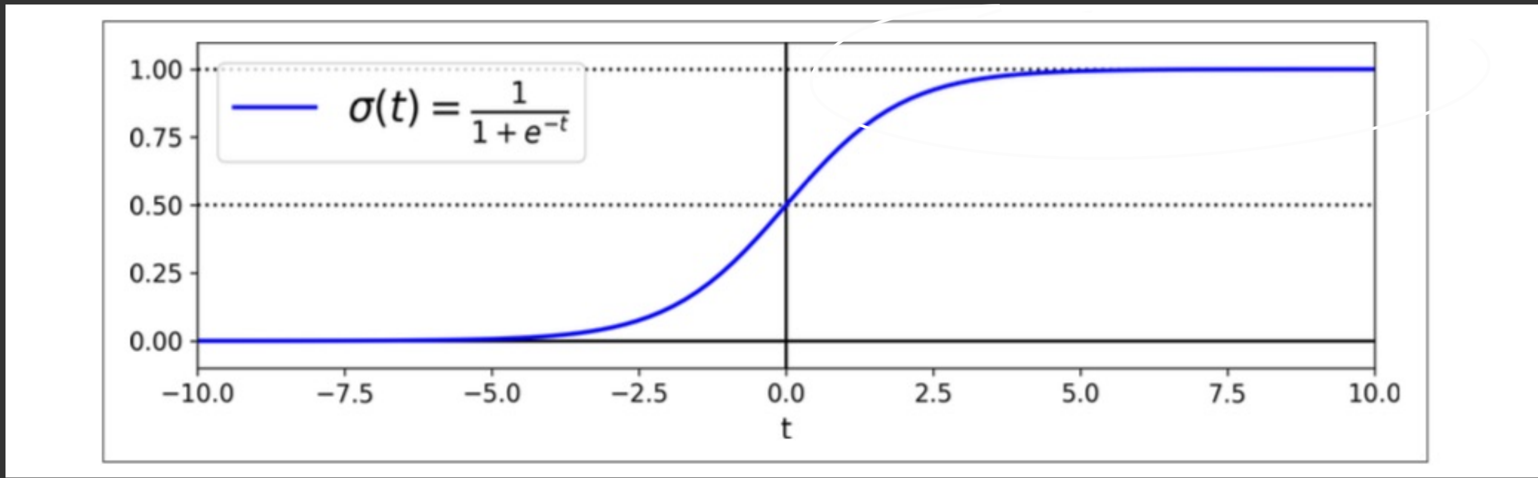
$$P(C_1 \mid \underline{x}) \propto P(\underline{x} \mid C_1)\, P(C_1)$$

$$P(C_1 \mid \underline{x}) = \cfrac{1}{1 + \cfrac{P(\underline{x} \mid C_2)\, P(C_2)}{P(\underline{x} \mid C_1)\, P(C_1)}}$$

$$= \frac{1}{1 + \exp(-t)} = \sigma(t)$$

$$t = \ln \frac{P(x \mid C_1)\, P(C_1)}{P(x \mid C_2)\, P(C_2)}$$

"logistic sigmoid"



<u>Properties:</u>

$$\sigma(-t) = 1 - \sigma(t)$$

$$t = \ln\left(\frac{\sigma}{1-\sigma}\right)$$

logits or log odds :

$$= \ln\left[\frac{P(C_1|\underline{x})}{P(C_2|\underline{x})}\right]$$

Multiclass: $(C > 2)$

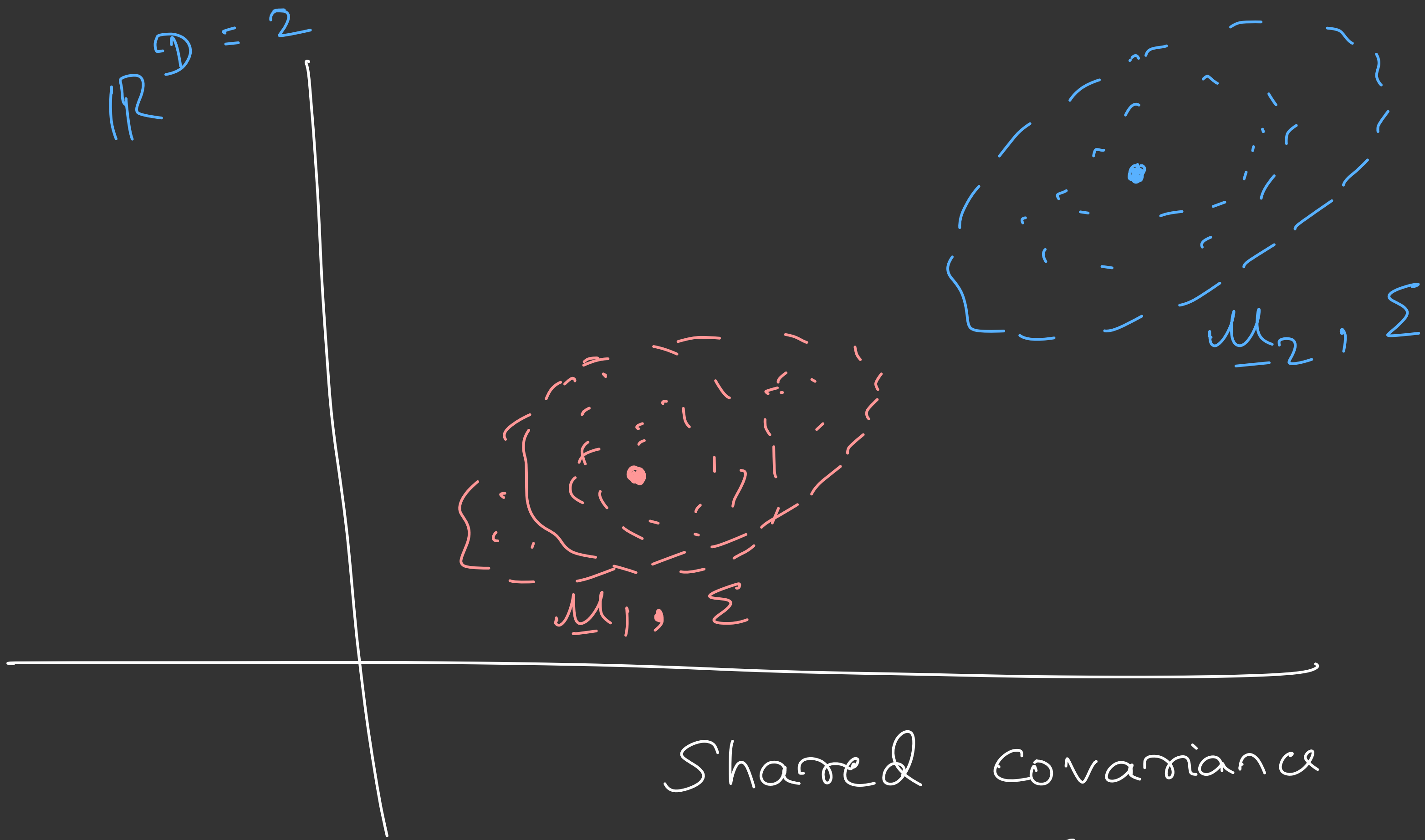
$$p(C_k|\underline{x}) = \frac{p(\underline{x}|C_k)\,p(C_k)}{\sum_j p(\underline{x}|C_j)\,p(C_j)} = \underbrace{\frac{\exp(a_k)}{\sum_j \exp(a_j)}}_{\text{Softmax or normalized exponential}}$$

with

$$a_k = \ln p(x|C_k)\,p(C_k)$$

Continuous input

Gaussian class-conditional densities:



Shared covariance matrix Σ

Assume

$$P(\underline{x} \mid C_k) = \mathcal{N}(\underbrace{\underline{\mu}_k}_{D\times 1}, \underbrace{\Sigma}_{D\times D})$$

$$P(\underline{x} \mid C_k) =$$

$$\frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} (\underline{x} - \underline{\mu}_k)^T \Sigma^{-1} (\underline{x} - \underline{\mu}_k)\right)$$

Compute t of $\sigma(t)$:

$$\frac{P(\underline{x} \mid C_2) \quad P(C_2)}{P(\underline{x} \mid C_1) \quad P(C_1)}$$

$$= \exp\left\{ -\frac{1}{2} (\underline{x} - \underline{\mu}_2)^T \Sigma^{-1} (\underline{x} - \underline{\mu}_2) + \frac{1}{2} (\underline{x} - \underline{\mu}_1)^T \Sigma^{-1} (\underline{x} - \underline{\mu}_1) + \ln \frac{P(C_2)}{P(C_1)} \right\}$$

$$t = \underline{w}^T \underline{x} + b$$

$$\underline{w} = \Sigma^{-1} (\underline{\mu}_1 - \underline{\mu}_2)$$

$$b = -\frac{1}{2} \underline{\mu}_1^T \Sigma^{-1} \underline{\mu}_1 + \frac{1}{2} \underline{\mu}_2^T \Sigma^{-1} \underline{\mu}_2 + \ln \frac{P(C_1)}{P(C_2)}$$

So $P(C_1 | \underline{x}) = \sigma(\underline{w}^T \underline{x} + b)$

Now, all that remains is to find $\underline{\mu}_1, \underline{\mu}_2, \Sigma$, and $p(c_1)$.
For which, we use the maximum likelihood estimation procedure

Maximum likelihood Estimation (MLE):

Define $p(c_1) = \pi$ and $p(c_2) = 1 - \pi$

For example i:

$$P(\underline{x}_i, c_1) = P(c_1)\, P(\underline{x}_i | c_1) = \pi\, \mathcal{N}(\underline{x}_i; \mu_1, \Sigma)$$

$$P(\underline{x}_i, c_2) = P(c_2)\, P(\underline{x}_i | c_2) = (1 - \pi)\, \mathcal{N}(\underline{x}_i; \mu_2, \Sigma)$$

For N examples in the training data:

$$\ln p\left(\{\underline{x}_i, y_i\}_{i=1}^{N} ; \pi, \mu_1, \mu_2, \Sigma\right)$$

$$= J\left(\pi, \mu_1, \mu_2, \Sigma\right)$$

$$= \ln \prod_{i=1}^{N} \left[\pi \, N(\underline{x}_i ; \underline{\mu}_1, \Sigma)\right]^{y_i} \left[(1-\pi) \, N(\underline{x}_i ; \underline{\mu}_2, \Sigma)\right]^{1-y_i}$$

maximize (log) likelihood wrt. $\pi, \underline{\mu}_1, \underline{\mu}_2, \Sigma$

① $$\frac{\partial J}{\partial \pi} = 0$$

$$\frac{\partial}{\partial \pi} \sum_{i=1}^{N} y_i \ln \pi + (1-y_i) \ln(1-\pi) = 0$$

$$\sum_{i=1}^{N} y_i \cdot \frac{1}{\pi} + (1-y_i)\frac{1}{1-\pi}(-1) = 0$$

$$P(c_1) = \pi = \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{N_1}{N_1+N_2}$$

(2) $$\frac{\partial J}{\partial \underline{\mu}_k} = 0 \quad \text{for } k = 1, 2$$

$$\frac{\partial}{\partial \underline{\mu}_k}\left[-\frac{1}{2}\sum_{i=1}^{N} y_i (\underline{x}_i - \underline{\mu}_k)^T \Sigma^{-1} (\underline{x}_i - \underline{\mu}_k) + \text{Const.}\right] = 0$$

$$\sum_{i=1}^{N} y_i \Sigma^{-1} \underline{\mu}_k = \sum_{i=1}^{N} y_i \Sigma^{-1} \underline{x}_i$$

$$\Rightarrow \quad \underline{\mu}_1 = \frac{1}{N_1} \sum_{i=1}^{N} y_i \underline{x}_i$$

$$\underline{\mu}_2 = \frac{1}{N_2} \sum_{i=1}^{N} (1 - y_i) \underline{x}_i$$

Sample mean of features belonging to class C_k

(3) $$\frac{\partial J}{\partial \Sigma} = 0$$

$$J = -\frac{1}{2} \sum_{i=1}^{N} y_i \ln |\Sigma| - \frac{1}{2} \sum_{i=1}^{N} y_i (\underline{x}_i - \underline{\mu}_1)^T \Sigma^{-1} (\underline{x}_i - \underline{\mu}_1)$$
$$- \frac{1}{2} \sum_{i=1}^{N} (1 - y_i) \ln |\Sigma| - \frac{1}{2} \sum_{i=1}^{N} (1 - y_i) \left[(\underline{x}_i - \underline{\mu}_2)^T \Sigma^{-1} (\underline{x}_i - \underline{\mu}_2) \right]$$

$$= -\frac{N}{2} \ln |\Sigma| - \frac{N}{2} \operatorname{tr}(\Sigma^{-1} S)$$

$$S = \frac{N_1}{N} S_1 + \frac{N_2}{N} S_2$$

$$S_1 = \frac{1}{N_1} \sum_{i \in C_1} (x_i - \mu_1)(x_i - \mu_1)^T$$

$$S_2 = \frac{1}{N_2} \sum_{i \in C_2} (x_i - \mu_2)(x_i - \mu_2)^T$$

Note : $\frac{\partial}{\partial \Sigma} \ln |\Sigma| = \Sigma^{-1}$ and $\frac{\partial}{\partial \Sigma} \operatorname{tr}(\Sigma^{-1} S) = -\Sigma^{-1} S \Sigma$

$$\Rightarrow \quad \frac{N}{2}\Sigma^{-1} - \frac{N}{2}\Sigma^{-1} S \Sigma = 0$$

$$I - S\Sigma^{-1} = 0$$

$$\boxed{\Sigma = S}$$

weighted sample covariance
(weighted by the fraction
of samples in each class)

Discrete Features $\underline{x} \in \{0,1\}^D$

General distribution would be a table with 2^D numbers per class (grows exponentially)

Naive Bayes Classifier:

$$P(\underline{x} \mid C_k) = \prod_{d=1}^{D} P(x_d \mid C_k)$$

Assume features are independent conditioned on the class

Coordinate d of the feature:

$$\text{Ber}(x_d \mid C_k ; \mu_{dk})$$

$$P(\underline{x} \mid C_k) = \prod_{d=1}^{D} \mu_{dk}^{x_d} (1-\mu_{dk})^{1-x_d}$$

Then,

$$\ln \frac{P(\underline{x} \mid C_1)\, P(C_1)}{P(\underline{x} \mid C_2)\, P(C_2)}$$

$$= \ln \prod_{d=1}^{D} \frac{\mu_{d1}^{x_d} (1-\mu_{d1})^{1-x_d}}{\mu_{d2}^{x_d} (1-\mu_{d2})^{1-x_d}} \cdot \frac{P(C_1)}{P(C_2)}$$

$$= \ln \frac{P(C_1)}{P(C_2)} + \sum_{d=1}^{D} x_d \ln \left[\frac{\mu_{d1}}{\mu_{d2}}\right] + (1-x_d) \ln \left[\frac{1-\mu_{d1}}{1-\mu_{d2}}\right]$$

$$= \underbrace{\ln \frac{P(C_1)}{P(C_2)} + \sum_{d=1}^{D} \ln \left[\frac{1-\mu_{d1}}{1-\mu_{d2}}\right]}_{b} + \sum_{d=1}^{D} x_d \underbrace{\left[\ln \left[\frac{\mu_{d1}}{\mu_{d2}}\right] - \ln \left[\frac{1-\mu_{d1}}{1-\mu_{d2}}\right]\right]}_{w}$$

$$P(c_1 \mid \underline{x}) = \sigma(\underline{w}^T \underline{x} + b)$$

$$[\underline{w}]_d = \ln\left(\frac{\mu_{d1}(1-\mu_{d2})}{\mu_{d2}(1-\mu_{d1})}\right)$$

$$b = \ln \frac{P(c_1)}{P(c_2)} + \sum_{d=1}^{D} \ln \left[\frac{1-\mu_{d1}}{1-\mu_{d2}}\right]$$

As before we can compute the parameters $\{\mu_{d1}\}$, $\{\mu_{d2}\}$, and π using MLE:

$$P\left(\{\underline{x}_i, y_i\}_{i=1}^{N} ; \{\mu_{kd}\}_{k=1, d=1}^{C, D}\right)$$

$$= \prod_{n=1}^{N} \left[\pi \prod_{d=1}^{D} \mu_{1d}^{x_d} (1-\mu_{1d})^{1-x_{nd}}\right]^{y_i} \left[(1-\pi) \prod_{d=1}^{D} \mu_{2d}^{x_d} (1-\mu_{2d})^{1-x_d}\right]^{1-y_i}$$

Take log & derive wrt π, $\{\mu_{kd}\}$ and set to zero:

$$\pi = \frac{1}{N} \sum_{i=1}^{N} y_i = \frac{N_1}{N} \quad ; \quad 1-\pi = \frac{N_2}{N}$$

$$\mu_{1d} = \sum_{i=1}^{N} y_i x_{id} \Big/ \sum_{i=1}^{N} y_i \quad \text{for } d = 1, \ldots, D$$

$$\mu_{2d} = \sum_{i=1}^{N} (1-y_i) x_d \Big/ \sum_{i=1}^{N} (1-y_i) \quad \text{for } d = 1, \ldots, D$$

Excercise:

Verify

Summary:

- For some data distributions (commonly used) we obtain the halfspace hypothesis class for which $\underline{w}$ admits a closed form expression that depends on the parameters of the data PDF.

- For Gaussian class-condition, we have to estimate: $\underbrace{D}_{\mu_1} + \underbrace{D}_{\mu_2} + \underbrace{\frac{D(D+1)}{2}}_{\Sigma} + \underbrace{1}_{\pi}$

$$= \frac{D(D+5)}{2} + 1 \quad \text{parameters}$$

- Naive Bayes: features are independent.