Lecture 16.1

$$h(\vec{x}) = \arg\max_y P_\theta(\vec{x}|y)\, P_\theta(y)$$

$$= \prod_{\alpha=1}^{d} P_\theta(x_\alpha|y) = \frac{m!}{x_1!x_2!\cdots x_d!} \prod_{\alpha=1}^{d} ([\theta_y]_\alpha)^{x_\alpha}$$

$y \in \{\text{spam} \mid \text{ham}\}$

$$\frac{P_\theta(x|y=\text{spam})\, P_\theta(y=\text{spam})}{P_\theta(x|y=\text{ham})\, P_\theta(y=\text{ham})} = \frac{\frac{m!}{x_1!x_2!\cdots x_d!}\left(\prod_{\alpha=1}^{d}([\theta_{\text{spam}}]_\alpha)^{x_\alpha}\right) P_\theta(y=\text{spam})}{\frac{m!}{x_1!x_2!\cdots x_d!}\left(\prod_{\alpha=1}^{d}([\theta_{\text{ham}}]_\alpha)^{x_\alpha}\right) P_\theta(y=\text{ham})}$$

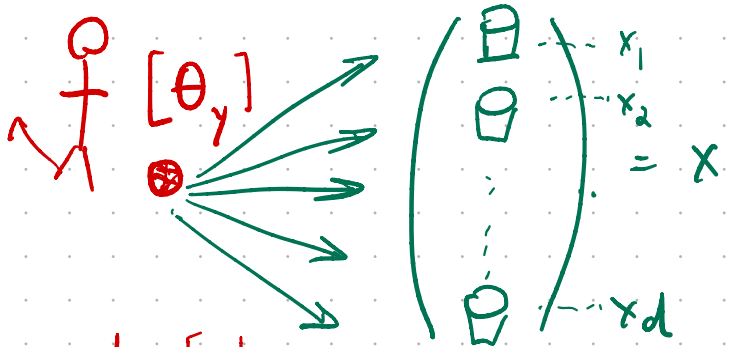
If the denominator $\geq$ the divisor, output "spam".
O.t.w., output "ham".

Alternatively
$$P_\theta(y|x) \propto \sum_{\alpha=1}^{d} x_\alpha \log([\theta_y]_\alpha) + \log(P_\theta(y))$$

Remark: The multinomial distribution is our modeling assumption for $P(x|y)$

$$P_\theta(x|y) = \frac{m!}{x_1!x_2!\cdots x_d!} \prod_{\alpha=1}^{d} ([\theta_y]_\alpha)^{x_\alpha}$$

multinomial distribution



Multinomial feature

x_α = # of times the ball goes to bin α
$\in \{0, 1, \ldots, m\}$
and $\sum_{\alpha=1}^{d} x_\alpha = m$

x_1, x_2, x_1, x_3, x_3

$$x = \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix}$$

x_a, x_2, x_2, x_3, x_3

Continuous Features (Gaussian Naïve Bayes)

$x_\alpha \in \mathbb{R}$ (e.g. heights, weights)

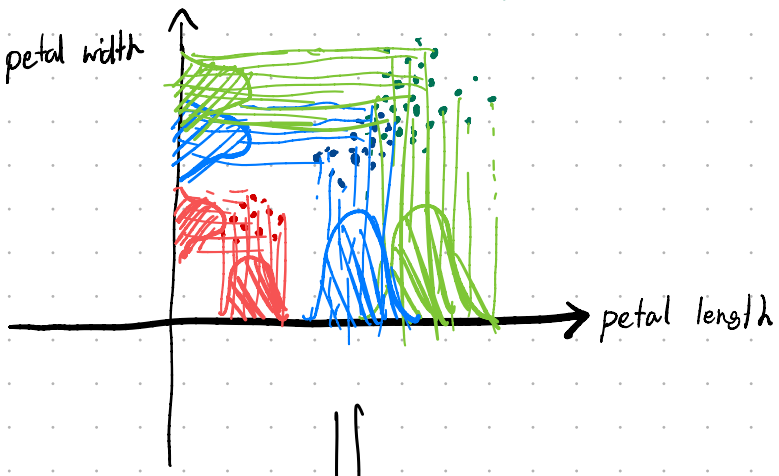
recall our task to classify fish type given the length

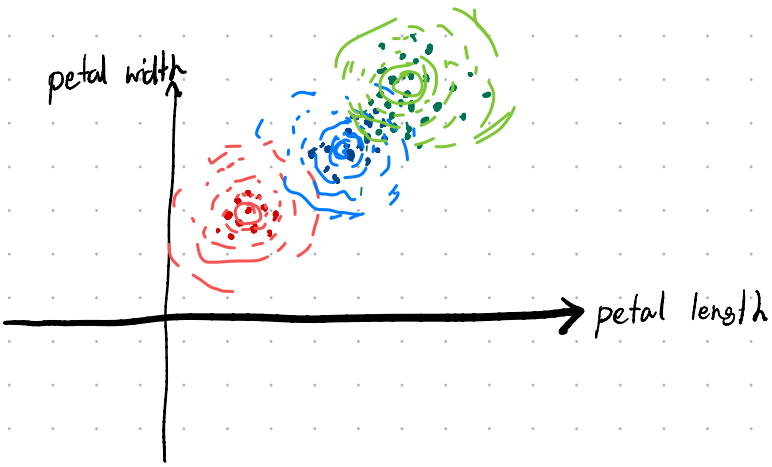
- Our modeling assumption:
  - $P_\theta(\vec{x}|y)$ = multivariate Gaussian
  - $P_{\theta_\alpha}(\vec{x}_\alpha|y)$ = univariate Gaussian with some mean and s.d.
    $= \mathcal{N}([\mu_y]_\alpha, [\sigma_y]_\alpha)$

$\Rightarrow \prod_{\alpha=1}^{d} P(\vec{x}_\alpha|y)$ is a multivariate Gaussian

(Gaussian × Gaussian → Gaussian)

Remark: We already know how to estimate $P(\vec{x}_\alpha|y)$





Summary of Naïve Bayes

- Naïve Bayes = Bayes classifier + naïve Bayes assumption
- The assumption says "all feature values are independent."
- We may have data that violates the assumptions.
- If your data is multinomial features for binary classification, then naïve Bayes always give the linear decision boundary (naïve Bayes is a linear classifier)

- Assume $y \in \{-1, +1\}$

$$h(\vec{x}) = +1 \quad \text{iff} \quad P(y=+1 \mid \vec{x}) > P(y=-1 \mid \vec{x})$$

$$\text{iff} \quad P(\vec{x} \mid y=+1)\,P(y=+1) > P(\vec{x} \mid y=-1)\cdot P(y=-1)$$

$$\text{iff} \quad \prod_{\alpha=1}^{d} P_\theta(x_\alpha \mid y=+1)\,P_\theta(y=+1) > \prod_{\alpha=1}^{d} P_\theta(x_\alpha \mid y=-1)\,P_\theta(y=-1)$$

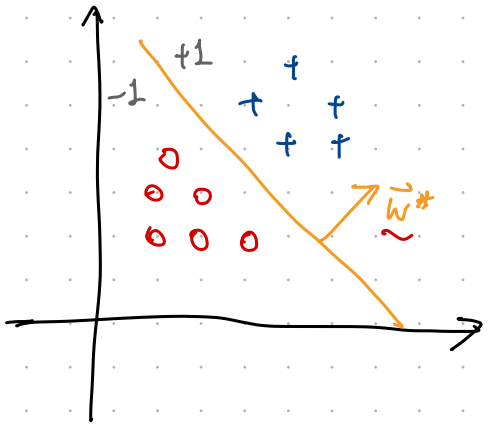
$$\text{iff} \quad \sum_{\alpha=1}^{d} x_\alpha \log\left(P_\theta(x_\alpha \mid y=+1)\right) + \log\left(P_\theta(y=+1)\right) > \sum_{\alpha=1}^{d} x_\alpha \log\left(P_\theta(x_\alpha \mid y=-1)\right) + \log\left(P_\theta(y=-1)\right)$$

$$\text{iff} \quad \sum_{\alpha=1}^{d} x_\alpha \left(\log\left(P_\theta(x_\alpha \mid y=+1)\right) - \log\left(P_\theta(x_\alpha \mid y=-1)\right)\right) + \left(\log\left(P_\theta(y=+1)\right) - \log\left(P_\theta(y=-1)\right)\right) > 0$$

$$\text{iff} \quad \sum_{\alpha=1}^{d} x_\alpha \underbrace{\left(\log\left([\theta_{y=+1}]_\alpha\right) - \log\left([\theta_{y=-1}]_\alpha\right)\right)}_{\vec{w}} + \underbrace{\left(\log\left(P(y=+1)\right) - \log\left(P(y=-1)\right)\right)}_{b} > 0$$

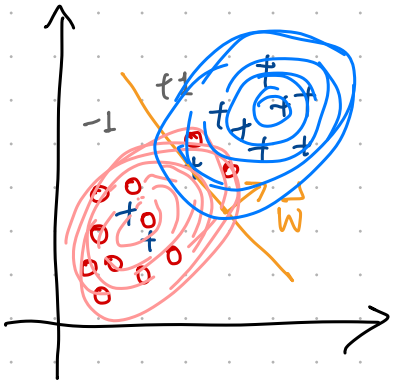
$$\text{iff} \quad w^T x + b > 0$$

w: hyperplane



Perceptron

Find w^* that best separates nagative points from positive points



Naïve Bayes + multinomial Features

Find $\vec{w}$ that best separates the two modeling distributions

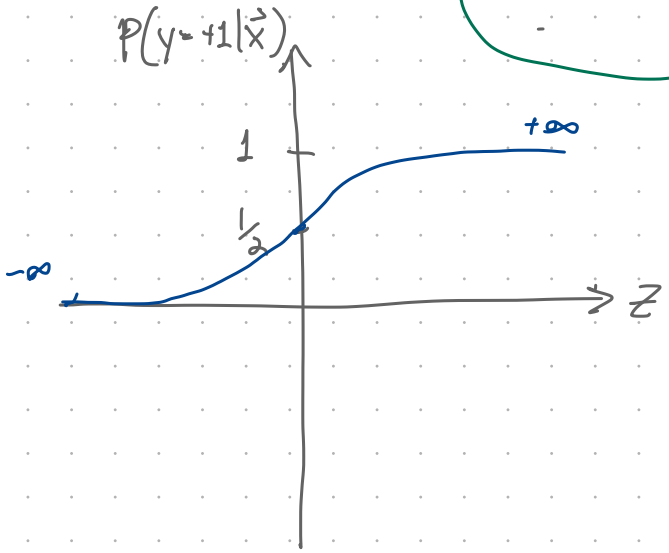
Discriminative algorithm: model $P(y|x)$ (e.g. k-NN, Perceptron)

Generative algorithm: model $P(x|y)$ and $P(y)$ to estimate $P(y|x)$ (e.g. naïve Bayes)

- For Gaussian naïve Bayes, you will arrive the following by taking the same derivation

$$P(y|\vec{x}) = \frac{1}{1+e^{-y(\vec{w}^T\vec{x}+b)}} \quad \text{for } y \in \{-1, +1\}$$

($z = \vec{w}^T\vec{x}+b$)



close-form solution for decision boundary

Here comes the Logistic Regression