CS 3780/5780
Logistic Regression

Recall Naive Bayes Model:

Assumption: $P(\vec{X} = \vec{x} \mid Y = y) = \prod_{\alpha=1}^{d} P(\vec{X}[\alpha] = \vec{x}[\alpha] \mid Y = y)$

$$\boxed{P(Y=y \mid X=x) = \frac{\prod_{\alpha=1}^{d} P(X[\alpha]=x[\alpha] \mid Y=y)\, P(Y=y)}{\sum_{c \in \mathcal{Y}} \prod_{\alpha=1}^{d} P(X[\alpha]=x[\alpha] \mid Y=c)\, P(Y=c)}}$$

Multinomial NB:

$$\boxed{P(X[\alpha]=x[\alpha] \mid Y=y) \propto \theta_y[\alpha]^{x[\alpha]}}$$

Eg $\mathcal{Y} = \{\text{spam}, \text{Not spam}\}$
$= \{+1, -1\}$

$\theta_{+1} =$ distribution of words in spam emails

Gaussian NB:

Same variance across class for each feature

$y = \text{Adult}$, μ_{Adult}

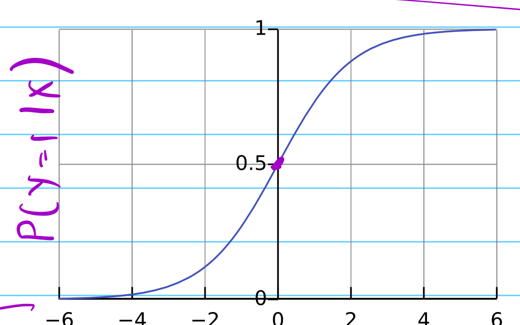
$$P(X[\alpha]=x[\alpha] \mid Y=y) = \frac{1}{\sqrt{2\pi\sigma_\alpha^2}} e^{-\frac{(x[\alpha]-\mu_y[\alpha])^2}{2\sigma_\alpha^2}}$$

Eg: Take $\mathcal{Y} = \{+1, -1\}$, in both the above cases:

show that $P(Y \mid X=x)$ has the following form:

Logit

$$P(Y=+1 \mid X=x) = \frac{1}{1 + e^{-(w^T x + b)}}$$



$w^T x + b$

In each of multinomial (and gaussian NB) cases, what are w and b?

Show for Multinomial NB case that:

$$P(y=1 \mid x=x) = \frac{P(x=x \mid y=1)\,P(y=1)}{P(x=x \mid y=1)\,P(y=1) + P(x=x \mid y=-1)\,P(y=-1)}$$

NB Assumption $=$

$$\frac{\prod_{\alpha=1}^{d} P(x[\alpha]=x[\alpha] \mid y=1)\,P(y=1)}{\prod_{\alpha=1}^{d} P(x[\alpha]=x[\alpha] \mid y=1)\,P(y=1) + \prod_{\alpha=1}^{d} P(x[\alpha]=x[\alpha] \mid y=-1)\,P(y=-1)}$$

$$= \frac{\prod_{\alpha=1}^{d} \left(\theta_{+1}[\alpha]\right)^{x[\alpha]}\,P(y=1)}{\prod_{\alpha=1}^{d} \left(\theta_{+1}[\alpha]\right)^{x[\alpha]}\,P(y=1) + \prod_{\alpha=1}^{d} \left(\theta_{-1}[\alpha]\right)^{x[\alpha]}\,P(y=-1)}$$

set $b_+ = \log(P(y=1))$

$w_{+1}[\alpha] = \log(\theta_{+1}[\alpha])$

$$= \frac{\prod_{\alpha=1}^{d} e^{w_+[\alpha].x[\alpha]} \times e^{b_+}}{\prod_{\alpha=1}^{d} e^{w_+[\alpha].x[\alpha]}\,e^{b_+} + \prod_{\alpha=1}^{d} e^{w_-[\alpha].x[\alpha]} \times e^{b_-}}$$

$$= \boxed{\frac{e^{w_{+1}^T x + b_+}}{e^{w_{+1}^T x + b_+} + e^{w_{-1}^T x + b_-}}}$$

$$= \frac{1}{1 + e^{w_-^T x - w_+^T x + b_- - b_+}}$$

$$= \frac{1}{1 + e^{-(w^T x + b)}}$$

$w = w_+ - w_-$ (in terms of w_+ and w_-)

$b = b_+ - b_-$ (in terms of b_+ and b_-)

Since $Y = \{+1, -1\}$ $\quad P(Y = y \mid \vec{X} = \vec{x}) = \dfrac{1}{1 + e^{-y(w^T x + b)}}$

NB is generative: we model $P(X, Y)$

Discriminative model we only model $P(Y \mid X)$

Discriminative counterpart of Multinomial NB (and Gaussian NB) is Logistic Regression.

Probabilistic model: (absorb bias into last dimension)

$$P(Y = \overset{+1}{y} \mid \vec{X} = \vec{x}) = \frac{1}{1 + e^{-y w^T x}}$$

$$
\begin{aligned}
\hat{\mathbf{w}}_{MLE} &= \operatorname*{argmax}_{\mathbf{w}} P(D|\mathbf{w}) && \text{(Definition of MLE)}\\
&= \operatorname*{argmax}_{\mathbf{w}} P((y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n) \mid \mathbf{w}) && \text{(Substituting in D.)}\\
&= \operatorname*{argmax}_{\mathbf{w}} \prod_{i=1}^{n} P(y_i, \mathbf{x}_i \mid \mathbf{w}) && \text{(Data is i.i.d.)}\\
&= \operatorname*{argmax}_{\mathbf{w}} \prod_{i=1}^{n} P(y_i \mid \mathbf{x}_i, \mathbf{w}) P(\mathbf{x}_i \mid \mathbf{w}) && \text{(Chain Rule of Statistics)}\\
&= \operatorname*{argmax}_{\mathbf{w}} \prod_{i=1}^{n} P(y_i \mid \mathbf{x}_i, \mathbf{w}) P(\mathbf{x}_i) && \text{(}\mathbf{x}_i \text{ does not depend on } \mathbf{w}\text{)}\\
&= \operatorname*{argmax}_{\mathbf{w}} \prod_{i=1}^{n} P(y_i \mid \mathbf{x}_i, \mathbf{w}) && \text{(}P(\mathbf{x}_i) \text{ does not affect } \mathbf{w}\text{)}\\
&= \operatorname*{argmax}_{\mathbf{w}} \sum_{i=1}^{n} \log\left[P(y_i \mid \mathbf{x}_i, \mathbf{w})\right]. && \text{(Taking the } \mathbf{log}\text{)}\\
&= \boxed{\operatorname*{argmax}_{\boxed{\mathbf{w}}} -\sum_{i=1}^{n} \log(1 + e^{-y_i \mathbf{w}^T \mathbf{x}_i})} && \text{(Substituting in } P(y_i \mid \mathbf{x}_i, \mathbf{w})\text{)}\\
&= \operatorname*{argmin}_{\mathbf{w}} \sum_{i=1}^{n} \log(1 + e^{-y_i \mathbf{w}^T \mathbf{x}_i}) && \text{(We prefer minimization.)}
\end{aligned}
$$

log

$\text{loss}(w, x_i, y_i)$

$\shortparallel$

$\log(1 + e^{-y_i w^T x_i})$

logistic loss

Find $\vec{w}$ s.t. $\boxed{\nabla \sum_{i=1}^{n} \log(1 + e^{-y_i w^T x_i}) = 0}$

No closed form, use GD to optimize

Maximum a posterior: Prior $w \sim N(0, \sigma^2 I)$

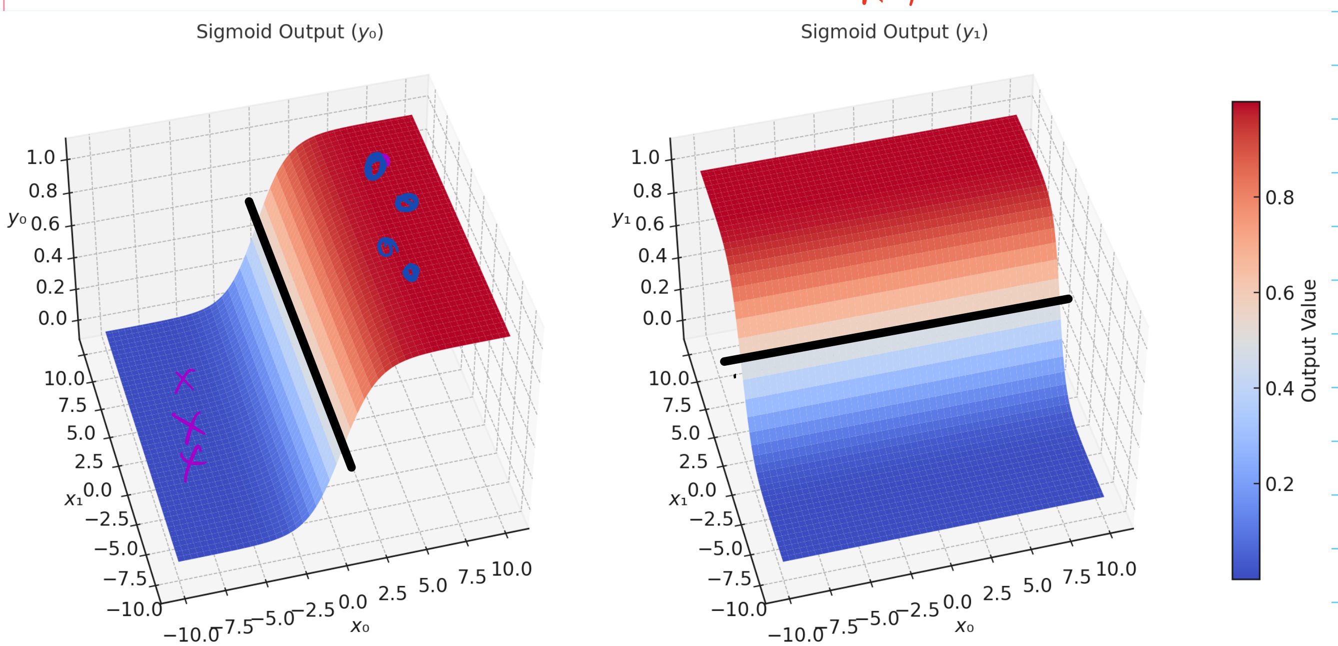
$$
\begin{aligned}
\hat{\mathbf{w}}_{MAP} &= \underset{\mathbf{w}}{\operatorname{argmax}}\, P(D|\mathbf{w})P(\mathbf{w}) \\
&= \underset{\mathbf{w}}{\operatorname{argmax}}\, P((y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n) \mid \mathbf{w})P(\mathbf{w}) \\
&= \underset{\mathbf{w}}{\operatorname{argmax}} \left(\prod_{i=1}^{n} P(y_i \mid \mathbf{x}_i, \mathbf{w}) \right) P(\mathbf{w}) \\
&= \underset{\mathbf{w}}{\operatorname{argmax}} \sum_{i=1}^{n} \log\left[P(y_i \mid \mathbf{x}_i, \mathbf{w})\right] + \boxed{\log P(\mathbf{w})} \\
&= \underset{\mathbf{w}}{\operatorname{argmin}} - \sum_{i=1}^{n} \log\left[P(y_i \mid \mathbf{x}_i, \mathbf{w})\right] - \log P(\mathbf{w}) \\
&= \underset{\mathbf{w}}{\operatorname{argmin}} \sum_{i=1}^{n} \log\left[1 + e^{-y_i \mathbf{w}^T \mathbf{x}_i}\right] + \frac{1}{2\sigma^2} \mathbf{w}^\top \mathbf{w} \\
&= \underset{\mathbf{w}}{\operatorname{argmin}} \sum_{i=1}^{n} \log\left[1 + e^{-y_i \mathbf{w}^T \mathbf{x}_i}\right] + \lambda \mathbf{w}^\top \mathbf{w}
\end{aligned}
$$

$\lambda = 1/2\sigma^2$

Multiclass version:

$y = [K]$ $\quad w_1, \ldots, w_K$

$$P(Y = y \mid X = x) = \frac{e^{w_y^T x}}{\sum_{k=1}^{K} e^{w_k^T x}}$$



Multinomial NB:

$$P(y=+1 \mid \vec{X}=\vec{x}) = \frac{\prod_{\alpha=1}^{d} P(x[\alpha]=x[\alpha] \mid y=+1)\; P(y=+1)}{\sum_{c\in\mathcal{Y}} \prod_{\alpha=1}^{d} P(x[\alpha]=x[\alpha] \mid y=c)\; P(y=c)}$$

$$= \frac{\prod_{\alpha=1}^{d} \theta_{\alpha,+1}^{x[\alpha]}\; P(y=+1)}{\prod_{\alpha=1}^{d} \theta_{\alpha,+1}^{x[\alpha]}\; P(y=+1) + \prod_{\alpha=1}^{d} \theta_{\alpha,-1}^{x[\alpha]}\; P(y=-1)}$$

pick $\vec{w}_y[\alpha] = \log \theta_{\alpha,y}$ and $b_y = \log(P(y=y))$

$$= \frac{\left(\prod_{\alpha=1}^{d} e^{\vec{w}_{+1}[\alpha]\,\vec{x}[\alpha]}\right) e^{b_{+1}}}{\left(\prod_{\alpha=1}^{d} e^{\vec{w}_{+1}[\alpha]\,\vec{x}[\alpha]}\right) e^{b_{+1}} + \prod_{\alpha=1}^{d} e^{\vec{w}_{-1}[\alpha]\,\vec{x}[\alpha]}\, e^{b_{-1}}}$$

$$= \frac{e^{\vec{w}_{+1}^{T}\vec{x} + b_{+1}}}{e^{\vec{w}_{+1}^{T}\vec{x} + b_{+1}} + e^{\vec{w}_{-1}^{T}x + b_{-1}}}$$

$$= \frac{1}{1 + e^{(\vec{w}_{-1} - \vec{w}_{+1})^{T}x + (b_{-1} - b_{+1})}}$$

$$\vec{w} = \vec{w}_{+1} - \vec{w}_{-1} \qquad b = b_{+1} - b_{-1}$$

$$= \frac{1}{1 + e^{-(w^{T}x + b)}}$$

$P(y = \text{spam})$ $\qquad$ $P(y = \text{N-spam})$

θ_{spam} = distribution over words when mail is spam

θ_{NS} = distribution over words when mail is not spam.

$$P(\vec{X} = \vec{x} \mid y = \text{s}) = \prod_{\alpha} P(X[\alpha] = x[\alpha] \mid y = \text{spam}) \propto (\theta_{spam}[\alpha])^{x[\alpha]}$$

d = # words in dictionary.

θ_{spam} = d dim vector

represents probability of each word in spam emails.

θ_{NS} = d dim vector

$$\boxed{P(\vec{X} = \vec{x} \mid y = \text{spam}) \propto \prod_{\alpha=1}^{d} (\theta_{spam}[\alpha])^{x[\alpha]}}$$

NB Assumption for multinomial:

Given Email is spam (or not spam)

There is a fixed distribution over words and each word is drawn independently of others.

X is a d dimensional vector
d = size of lexicon

X[3] = # occurences of third word in dictionary.

$$P(X \mid Y = spam) \propto \prod_{\alpha=1}^{d} \left(\theta_{spam}[\alpha]\right)^{x[\alpha]}$$

"Multinomial distribution"

$$= \frac{m!}{x[1]! \cdots x[d]!} \prod_{\alpha=1}^{d} \left(\theta_{spam}[\alpha]\right)^{x[\alpha]}$$