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:

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

Gaussian NB:

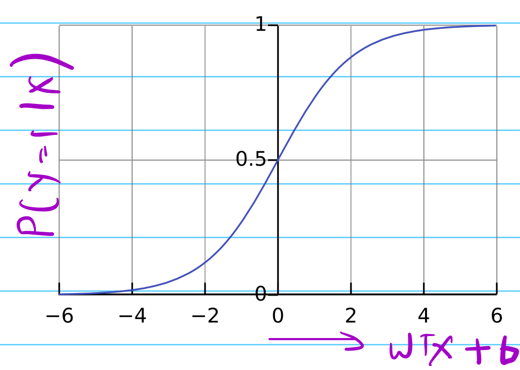
Same variance across class for each feature

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

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

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

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



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

Show for Multinomial NB case that:

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

$=$

$=$

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

$=$

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

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

$b =$ (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 $P(Y|X)$

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

Probabilistic model: (absorb bias into last dimension)

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

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

Find $\vec{w}$ st.

$$\nabla \sum_{i=1}^{n} \log(1+e^{-y_i w^T x_i}) = 0$$

No closed form, use GD to optimize

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

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

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

Multiclass version: $y = [K]$ $\quad w_1 - - w_K$

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