This is the discrimitave counterpart to Naive Bayes.
Here, we model $\mathbb{P}(y|\vec{x}_i)$ and assumes it takes exactly this form:

$$\mathbb{P}(y|\vec{x}_i) = \frac{1}{1 + e^{-(w^T x_i + b)}}$$

We make little assumptions on $\mathbb{P}(\vec{x}_i|y)$, because we we estimate vector $\vec{w}$ and b directly w/ MLE or MAP to maximize the conditional likelihood of

$$\prod_i \mathbb{P}(y_i|x_i; \vec{w}, b)$$

$\bar{x}_i \leftarrow \begin{bmatrix} x_i \\ 1 \end{bmatrix}$

$\bar{w} \leftarrow \begin{bmatrix} w \\ b \end{bmatrix}$

(makes math easier)

Maximum Likelihood Estimate

Want parameters that **maximize** the conditional likelihood.
The conditional data likelihood $\mathbb{P}(\vec{y}|X, \bar{w})$ is the probability of the observed values $\vec{y} \in \mathbb{R}^n$ in the training data conditioned on the feature values $\vec{x}_i$.
Note that

$$X = [\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n] \in \mathbb{R}^{(d+1) \times n}$$

We choose the parameters that maximize this function and we assume that the y_i's are independent given the input features $\bar{x}_i$ and $\bar{w}$.
So,

$$\mathbb{P}(y|X, \bar{w}) = \prod_{i=1}^{n} \mathbb{P}(y_i|\bar{x}_i, \bar{w})$$

$\bar{w}^T \bar{x}_i = wx_i + b$

Now if we take the log, we obtain

$$\log\left(\prod_{i=1}^{n} P(y_i | \bar{x}_i, \bar{w})\right) = -\sum_{i=1}^{n} \log\left(1 + e^{-y_i \bar{w}^T \bar{x}_i}\right)$$

$$\hat{\bar{w}} = \underset{w}{\operatorname{argmax}}\left(-\sum_{i=1}^{n} \log\left(1 + e^{-y_i \bar{w}^T \bar{x}_i}\right)\right)$$

$$= \underset{w}{\operatorname{argmin}} \sum_{i=1}^{n} \log\left(1 + e^{-y_i \bar{w}^T \bar{x}_i}\right)$$

convex function!

We need to estimate the parameters $\bar{w}$.

To find these values of the parameters at minimum, we can try to find solutions for

$$\nabla_{\bar{w}} \sum_{i=1}^{n} \log\left(1 + e^{-y_i \bar{w}^T \bar{x}_i}\right) = 0$$

This equation has **NO** closed form solution, so we use gradient descent (next lecture) on the negative log likelihood

$$\ell(\bar{w}) = \sum_{i=1}^{n} \log\left(1 + e^{-y_i \bar{w}^T \bar{x}_i}\right)$$

Maximum a Posteriori Estimate

In **MAP** estimate we treat $\bar{w}$ as a random variable and can specify a prior belief distribution over it.

We may use:

$$\bar{w} \sim \mathcal{N}(0, \sigma^2 I)$$

← Gaussian approximation for Logistic Regression

Our goal in MAP is to find the most likely model parameters given the data (i.e parameters that maximize the posterior)

$$P(\bar{w} \mid D) = P(\bar{w} \mid X, \vec{y}) \propto P(\vec{y} \mid X, \bar{w}) P(\bar{w})$$

$$\hat{w}_{MAP} = \underset{w}{\operatorname{argmax}} \log\left(P(\vec{y} \mid X, \bar{w}) P(\bar{w})\right)$$

$$= \underset{w}{\operatorname{argmin}} \sum_{i=1}^{n} \log\left(1 + e^{-y_i \bar{w}^T \bar{x}_i}\right) + \lambda \bar{w}^T \bar{w}$$

where $\lambda = \frac{1}{2\sigma^2}$

This function has no closed form solution, but we can use gradient descent on the negative log likelihood

$$\ell(\bar{w}) = \sum_{i=1}^{n} \log\left(1 + e^{-y_i \bar{w}^T \bar{x}_i}\right) + \lambda \bar{w}^T \bar{w}$$

to find optimal parameters

Summary

In Naive Bayes, we first model $\mathbb{P}(\vec{x}|y)$ for each label y, and then obtain the decision boundary that best discriminates between these two distributions.

In logistic regression, we do **NOT** attempt to model the data distribution $\mathbb{P}(\vec{x}|y)$, instead, we mode $\mathbb{P}(y|\vec{x})$ directly.
We assume the same probabilistic form $\mathbb{P}(y|\vec{x}_i) = 1/(1+e^{-y_i(\vec{w}^T\vec{x}_i+b)})$, but we do **NOT** restrict ourselves by making any assumptions about $\mathbb{P}(\vec{x}|y)$ (in fact it can be a member of any exponential family).
This allows logistic regression to be more flexible, but the flexibility comes at a cost - **NEED** more data to avoid overfitting.

Typically, in scenarios w/ little data, and if the modeling assumption is appropriate, Naive Bayes tends to **OUTPERFORM** logistic regression

However, as datasets become large logistic regression often outperforms Naive Bayes, which suffers from the fact that the assumptions made on $\mathbb{P}(\vec{x}|y)$ are probably not exactly correct

If assumptions hold exactly, i.e the data is truly drawn from the distribution that we assumed in Naive Bayes, then Logistic Regression and Naive Bayes converge to the exact same result in the limit (but NB will be faster)