

Probabilistic Learning and Logistic Regression

EECS 349 Spring 2016

Learning parameters in Bayes Nets

- ▶ Just statistical estimation for each parameter

Training Data

A	B
1	1
1	0
1	0
0	1
1	1
0	1
1	1

$$P_{ML}(A=1) = 5/7$$

$$P_{ML}(B=1, A=1) = 3/7$$



Discriminative vs. Generative training

- ▶ Say our distribution has variables $\mathbf{X}$, $\mathbf{Y}$
- ▶ Standard learning learns $P(\mathbf{X}, \mathbf{Y})$
- ▶ But often, the only inferences we care about are of form $P(\mathbf{Y} | \mathbf{X})$
 - ▶ $P(\text{Disease} | \text{Symptoms} = \mathbf{e})$
 - ▶ $P(\text{StockMarketCrash} | \text{RecentPriceActivity} = \mathbf{e})$



Discriminative vs. Generative training

- ▶ Learning $P(\mathbf{X}, \mathbf{Y})$: **generative** training
 - ▶ Learned model can “generate” the full data $\mathbf{X}, \mathbf{Y}$
- ▶ Learning only $P(\mathbf{Y} | \mathbf{X})$: **discriminative** training
 - ▶ Model **can't** assign probs. to $\mathbf{X}$. Only $\mathbf{Y}$ given $\mathbf{X}$
- ▶ Idea: Only model what we care about
 - ▶ Don't “waste data” on params irrelevant to task
 - ▶ Side-step false independence assumptions in training (example to follow)



Generative Model Example

- ▶ **Naïve Bayes model**

- ▶ Y binary $\{1=\text{spam}, 0=\text{not spam}\}$
 $\mathbf{X}$ an n -vector: message has word (1) or not (0)
- ▶ Re-write $P(Y | \mathbf{X})$ using Bayes Rule, apply Naïve Bayes assumption
- ▶ $2n + 1$ parameters, for n observed variables



Generative => Discriminative (1 of 3)

- ▶ But $P(Y | \mathbf{X})$ can be written more compactly

$$P(Y | \mathbf{X}) = \frac{1}{1 + \exp(w_0 + w_1 x_1 + \dots + w_n x_n)}$$

- ▶ Total of $n + 1$ parameters w_i



Generative => Discriminative (2 of 3)

- ▶ One way to do conversion (vars binary):

$$\exp(w_0) = \frac{P(Y = 0) P(X_1=0|Y=0) P(X_2=0|Y=0) \dots}{P(Y = 1) P(X_1=0|Y=1) P(X_2=0|Y=1) \dots}$$

for $i > 0$:

$$\exp(w_i) = \frac{P(X_i=0|Y=1) P(X_i=1|Y=0)}{P(X_i=0|Y=0) P(X_i=1|Y=1)}$$



Generative => Discriminative (3 of 3)

- ▶ We reduced $2n + 1$ parameters to $n + 1$
 - ▶ This must be *better*, right?
- ▶ Not exactly. If we construct $P(Y | \mathbf{X})$ to be equivalent to Naïve Bayes (as on prev. slide)
 - ▶ then it's...equivalent to Naïve Bayes
- ▶ Idea: optimize the $n + 1$ parameters directly, using training data



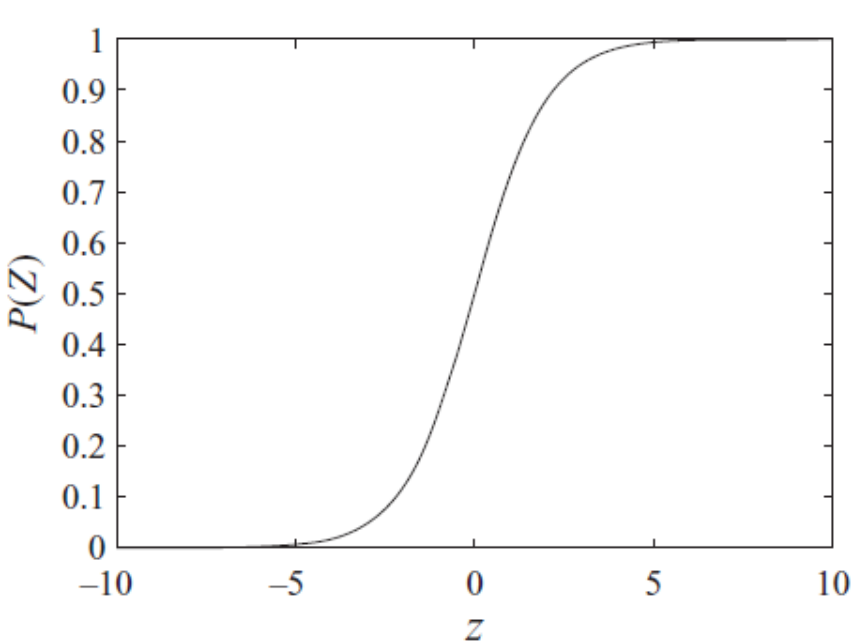
Discriminative Training

- ▶ In our example:

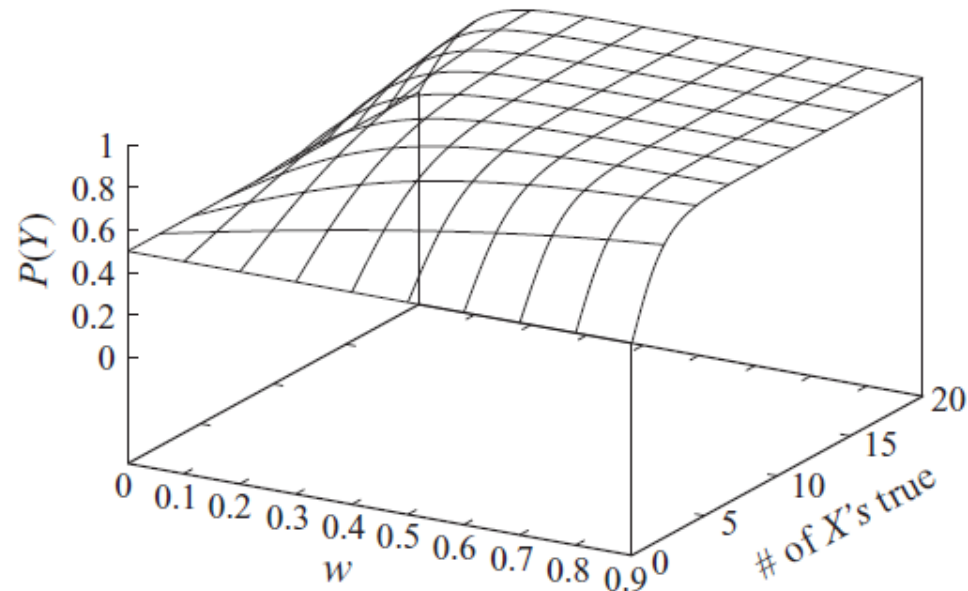
$$P(Y | \mathbf{X}) = \frac{1}{1 + \exp(w_0 + w_1 x_1 + \dots + w_n x_n)}$$

- ▶ Goal: find w_i that maximize likelihood of training data Y s given training data $\mathbf{X}$ s
 - ▶ Known as “logistic regression”
 - ▶ Solved with gradient ascent techniques
 - ▶ A convex optimization problem

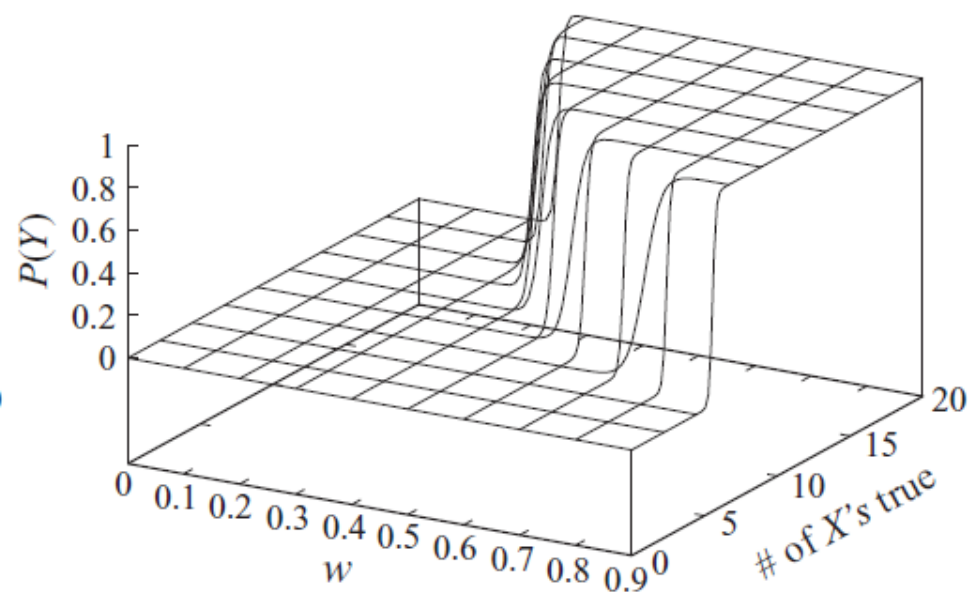
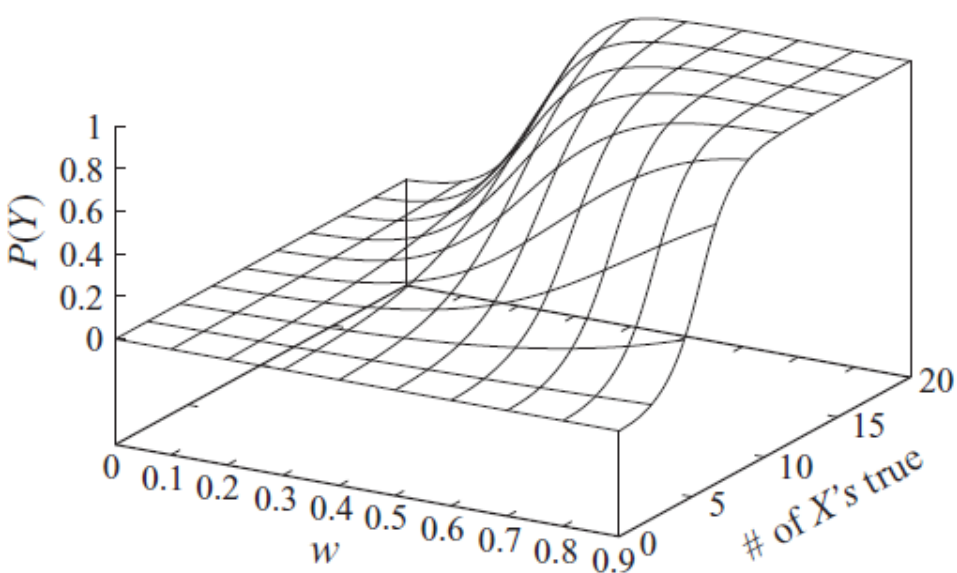




(a)



(b)



Naïve Bayes vs. LR

- ▶ Naïve Bayes “trusts its assumptions” in training
- ▶ Logistic Regression doesn’t – recovers better when assumptions violated



NB vs. LR: Example

Training Data

SPAM	Lottery	Winner	Lunch	Noon
1	1	1	0	0
1	1	1	1	1
0	0	0	1	1
0	1	1	0	1

- ▶ Naïve Bayes will classify the last example incorrectly, even after training on it!
- ▶ Whereas Logistic Regression is perfect with e.g.,
 $w_0 = 0.1 \quad w_{\text{lottery}} = w_{\text{winner}} = w_{\text{lunch}} = -0.2 \quad w_{\text{noon}} = 0.4$



Logistic Regression in practice

- ▶ Can be employed for any numeric variables X_i
 - ▶ or for other variable types, by converting to numeric (e.g. indicator) functions
- ▶ “Regularization” plays the role of priors in Naïve Bayes
- ▶ Optimization tractable, but (way) more expensive than counting (as in Naïve Bayes)



Discriminative Training

- ▶ Naïve Bayes vs. Logistic Regression one illustrative case
- ▶ Applicable more broadly, whenever queries $P(\mathbf{Y} \mid \mathbf{X})$ known *a priori*

