

## Generative and Discriminative

Logistic regression models only the conditional probability of labels  $p(y|\mathbf{x})$  and not the full data  $p(\mathbf{x}, y)$ . We can factorize the latter as  $p(y)p(\mathbf{x}|y)$  in a *generative* fashion.

$$p(\text{data}) = \prod_n p(y_n)p(\mathbf{x}_n | y_n)$$

For binary,  $p(y)$  can be bernoulli; for multiclass, multinomial. In either case, parameterize by  $\pi$ .  $p(\mathbf{x}|y)$  is often more complicated. Suppose there are  $F$ -many features; then:

$$p(\mathbf{x} | y) = p(x_1 | y)p(x_2 | y, x_1) \cdots p(x_F | y, \mathbf{x}_{1:F-1}) = \prod_f p(x_f | y, \mathbf{x}_{1:f-1})$$

We approximate this with the *naive Bayes* assumption of feature independence (conditional on  $y$ ):

$$p(\mathbf{x} | y) = \prod_f p(x_f | y, \mathbf{x}_{1:f-1}) \approx \prod_f p(x_f | y)$$

When features are binary, make each  $p(x_f | y)$  another bernoulli:

$$p(x_f | y; \theta) = \theta_f^{x_f} (1 - \theta_f)^{1-x_f}$$

Putting this together:

$$p(\text{data}) = \prod_n \pi^{y_n} (1 - \pi)^{1-y_n} \prod_f \theta_{y_n, f}^{x_{n, f}} (1 - \theta_{y_n, f})^{1-x_{n, f}}$$

**Exercise:** verify that there exist parameters  $\pi, \theta$  for the naive Bayes model that mimick the logistic regression behavior (and vice-versa).

Now, we want to do *maximum likelihood* estimation in this model. Consider our usual data set:

Class	Outlook	Temperature	Windy?
Play	Sunny	Low	Yes
Play	Overcast	Low	Yes
Play	Overcast	High	No
Play	Overcast	Low	No
Play	Rainy	Low	No
No play	Sunny	High	Yes
No play	Sunny	High	No
No play	Rainy	Low	Yes

The “obvious” estimate are just relative frequencies:

$p(y = \text{Play})$	$= \frac{5}{8}$
$p(y = \text{NoPlay})$	$= \frac{3}{8}$
$p(\text{Sunny} \mid \text{Play})$	$= \frac{1}{5}$
$p(\text{Overcast} \mid \text{Play})$	$= \frac{3}{5}$
$p(\text{Rainy} \mid \text{Play})$	$= \frac{1}{5}$
$p(\text{Low} \mid \text{Play})$	$= \frac{4}{5}$
$p(\text{High} \mid \text{Play})$	$= \frac{1}{5}$
$p(\text{Yes} \mid \text{Play})$	$= \frac{2}{5}$
$p(\text{No} \mid \text{Play})$	$= \frac{3}{5}$
$p(\text{Sunny} \mid \text{NoPlay})$	$= \frac{2}{3}$
$p(\text{Overcast} \mid \text{NoPlay})$	$= \frac{0}{3}$
$p(\text{Rainy} \mid \text{NoPlay})$	$= \frac{1}{3}$
$p(\text{Low} \mid \text{NoPlay})$	$= \frac{1}{3}$
$p(\text{High} \mid \text{NoPlay})$	$= \frac{2}{3}$
$p(\text{Yes} \mid \text{NoPlay})$	$= \frac{2}{3}$
$p(\text{No} \mid \text{NoPlay})$	$= \frac{1}{3}$

These can be derived directly from maximum likelihood; the tricky (and fun) example is the multinomial case, where we have to use the trick of Lagrange multipliers to derive the result.