Discriminative vs. Generative Learning

CS 760@UW-Madison
you should understand the following concepts

• the relationship between logistic regression and Naïve Bayes
• the relationship between discriminative and generative learning
• when discriminative/generative is likely to learn more accurate models
Discriminative vs. Generative

Discriminative approach:
• hypothesis $h \in H$ directly predicts the label given the features
  \[ y = h(x) \text{ or more generally, } p(y|x) = h(x) \]
• then define a loss function $L(h)$ and find hypothesis with min. loss

Generative approach:
• hypothesis $h \in H$ specifies a generative story for how the data was created:
  \[ p(x, y) = h(x, y) \]
• then pick a hypothesis by maximum likelihood estimation (MLE) or Maximum A Posteriori (MAP)
Summary: generative approach

- Step 1: specify the joint data distribution (generative story)
- Step 2: use MLE or MAP for training
- Step 3: use Bayes’ rule for inference on test instances

- Example: Naïve Bayes (conditional independence)

\[
p(x, y) = p(y)p(x|y) = p(y) \prod_{i} p(x_i|y)
\]
Summary: discriminative approach

- Step 1: specify the hypothesis class
- Step 2: specify the loss
- Step 3: design optimization algorithm for training

How to design the hypotheses and the loss? Can design by a generative approach!

- Step 0: specify $p(x|y)$ and $p(y)$
- Step 1: compute hypotheses $p(y|x)$ using Bayes’ rule
- Step 2: use conditional MLE to derive the negative log-likelihood loss (or use MAP to derive the loss)
- Step 3: design optimization algorithm for training

- Example: logistic regression
Logistic regression

• Suppose the class-conditional densities $p(x|y)$ is normal

$$p(x|y) = p(x|Y = y) = N(x|\mu_y, I) = \frac{1}{(2\pi)^{d/2}} \exp \left\{ -\frac{1}{2} \|x - \mu_y\|^2 \right\}$$

• Then conditional probability by Bayes’ rule:

$$p(Y = y|x) = \frac{p(x|Y = y)p(Y = y)}{\sum_k p(x|Y = k)p(Y = k)} = \frac{\exp(a_y)}{\sum_k \exp(a_k)}$$

where

$$a_k := \ln [p(x|Y = k)p(Y = k)] = -\frac{1}{2} x^T x + (w^k)^T x + b^k$$

with

$$w^k = \mu_k, \quad b^k = -\frac{1}{2} \mu_k^T \mu_k + \ln p(Y = k) + \ln \frac{1}{(2\pi)^{d/2}}$$
Logistic regression

• Suppose the class-conditional densities $p(x|y)$ is normal

$$ p(x|y) = p(x|Y = y) = N(x|\mu_y, I) = \frac{1}{(2\pi)^{d/2}} \exp \left\{ -\frac{1}{2} \|x - \mu_y\|^2 \right\} $$

• Cancel out $-\frac{1}{2} x^T x$, we have

$$ p(Y = y|x) = \frac{\exp(a_y)}{\sum_k \exp(a_k)}, \quad a_k := (w^k)^T x + b^k $$

where

$$ w^k = \mu_k, \quad b^k = -\frac{1}{2} \mu_k^T \mu_k + \ln p(Y = k) + \ln \frac{1}{(2\pi)^{d/2}} $$
Logistic regression: summary

- Suppose the class-conditional densities $p(x|y)$ is normal
  
  $$p(x|y) = p(x|Y = y) = N(x|\mu_y, I) = \frac{1}{(2\pi)^{d/2}} \exp \left\{ -\frac{1}{2} ||x - \mu_y||^2 \right\}$$

- Then
  
  $$p(Y = y|x) = \frac{\exp((w^y)^T x + b^y)}{\sum_k \exp((w^k)^T x + b^k)}$$

  which is the hypothesis class for multiclass logistic regression

- Training: find parameters $\{w^k, b^k\}$ that minimize the negative log-likelihood loss
  
  $$-\frac{1}{m} \sum_{j=1}^{m} \log p(y = y^{(j)}|x^{(j)})$$
Naïve Bayes vs. Logistic Regression
• Interesting observation: logistic regression is derived from the generative story

\[ p(x|y) = p(x|Y = y) = N(x|\mu_y, I) = \frac{1}{(2\pi)^{d/2}} \exp \left\{ -\frac{1}{2} \|x - \mu_y\|^2 \right\} \]

\[ = \frac{1}{(2\pi)^{d/2}} \prod_i \exp \left\{ -\frac{1}{2} (x_i - u_{yi})^2 \right\} \]

which is a special case of Naïve Bayes!

• Is the general Naïve Bayes assumption enough to get logistic regression? (Instead of the more special Normal distribution assumption)

• Yes, with an additional linearity assumption
Naïve Bayes revisited

consider Naïve Bayes for a binary classification task

\[
P(Y = 1 \mid x_1, \ldots, x_n) = \frac{P(Y = 1) \prod_{i=1}^{n} P(x_i \mid Y = 1)}{P(x_1, \ldots, x_n)}
\]

expanding denominator

\[
P(Y = 1) \prod_{i=1}^{n} P(x_i \mid Y = 1) = \frac{P(Y = 1) \prod_{i=1}^{n} P(x_i \mid Y = 1)}{P(Y = 1) \prod_{i=1}^{n} P(x_i \mid Y = 1) + P(Y = 0) \prod_{i=1}^{n} P(x_i \mid Y = 0)}
\]

dividing everything by numerator

\[
= \frac{1}{P(Y = 0) \prod_{i=1}^{n} P(x_i \mid Y = 0)}
\]

\[1 + \frac{1}{P(Y = 1) \prod_{i=1}^{n} P(x_i \mid Y = 1)}
\]
Naïve Bayes revisited

\[ P(Y = 1 | x_1, \ldots, x_n) = \frac{1}{P(Y = 0) \prod_{i=1}^{n} P(x_i | Y = 0)} \]

\[ + 1 + \frac{P(Y = 1) \prod_{i=1}^{n} P(x_i | Y = 1)}{P(Y = 1) \prod_{i=1}^{n} P(x_i | Y = 1)} \]

applying \( \exp(\ln(a)) = a \)

\[ = \frac{1}{\exp(\ln(1 + \prod_{i=1}^{n} P(x_i | Y = 0) \prod_{i=1}^{n} P(x_i | Y = 1)))} \]

applying \( \ln(a/b) = -\ln(b/a) \)

\[ = \frac{1}{\exp(-\ln\left(\frac{P(Y = 1) \prod_{i=1}^{n} P(x_i | Y = 1)}{P(Y = 0) \prod_{i=1}^{n} P(x_i | Y = 0)}\right))} \]
Naïve Bayes revisited

\[ P(Y = 1 \mid x_1, \ldots, x_n) = \frac{1}{1 + \exp \left( - \ln \left( \frac{P(Y = 1) \prod_{i=1}^{n} P(x_i \mid Y = 1)}{P(Y = 0) \prod_{i=1}^{n} P(x_i \mid Y = 0)} \right) \right)} \]

converting log of products to sum of logs

\[ P(Y = 1 \mid x_1, \ldots, x_n) = \frac{1}{1 + \exp \left( - \ln \left( \frac{P(Y = 1)}{P(Y = 0)} \right) - \sum_{i=1}^{n} \ln \left( \frac{P(x_i \mid Y = 1)}{P(x_i \mid Y = 0)} \right) \right)} \]

Does this look familiar?
Naïve Bayes vs. logistic regression

**Naïve Bayes**

\[ P(Y = 1 \mid x_1, \ldots, x_n) = \frac{1}{1 + \exp \left( - \ln \left( \frac{P(Y = 1)}{P(Y = 0)} \right) - \sum_{i=1}^{n} \ln \left( \frac{P(x_i \mid Y = 1)}{P(x_i \mid Y = 0)} \right) \right) } \]

**logistic regression**

\[ f(x) = \frac{1}{1 + \exp \left( - \left( w_0 + \sum_{i=1}^{n} w_i x_i \right) \right) } \]

Linearity assumption: the log-ratio is linear in \( x \).
Naïve Bayes vs. logistic regression

Naïve Bayes

\[
P(Y = 1 \mid x_1, \ldots, x_n) = \frac{1}{1 + \exp \left( -\ln \left( \frac{P(Y = 1)}{P(Y = 0)} \right) - \sum_{i=1}^{n} \ln \left( \frac{P(x_i \mid Y = 1)}{P(x_i \mid Y = 0)} \right) \right)}
\]

logistic regression

\[
f(x) = \frac{1}{1 + \exp \left( - \left( w_0 + \sum_{i=1}^{n} w_i x_i \right) \right)}
\]

Linearity assumption: the log-ratio is linear in \( x \)

Summary: If we begin with a Naïve Bayes generative story to derive a discriminative approach (assuming linearity), we get logistic regression!
Naïve Bayes vs. logistic regression

Naïve Bayes

\[ P(Y = 1 \mid x_1, \ldots, x_n) = \frac{1}{1 + \exp \left( -\ln \left( \frac{P(Y = 1)}{P(Y = 0)} \right) - \sum_{i=1}^{n} \ln \left( \frac{P(x_i \mid Y = 1)}{P(x_i \mid Y = 0)} \right) \right) } \]

logistic regression

\[ f(x) = \frac{1}{1 + \exp \left( - \left( w_0 + \sum_{i=1}^{n} w_i x_i \right) \right) } \]

Summary: If we begin with a Naïve Bayes generative story to derive a discriminative approach (assuming linearity), we get logistic regression!
Naïve Bayes vs. logistic regression

Conditional Independence (Naïve Bayes assumption)

Generative approach
- Naïve Bayes method

Discriminative approach (+ linearity assumption)
- Logistic regression
Logistic regression as a neural net

The connection can give interpretation for the weights in logistic regression: weights correspond to log ratios
Which is better?
they have the same functional form, and thus have the same hypothesis space bias (recall our discussion of inductive bias)

• Do they learn the same models?

In general, no. They use different methods to estimate the model parameters.

Naïve Bayes uses MLE to learn the parameters $p(x_i|y)$, whereas LR minimizes the loss to learn the parameters $w_i$. 
Naïve Bayes vs. logistic regression

asymptotic comparison ($\# \text{ training instances} \to \infty$)

• when conditional independence assumptions made by NB are correct, NB and LR produce identical classifiers

when conditional independence assumptions are incorrect

• logistic regression is less biased; learned weights may be able to compensate for incorrect assumptions (e.g. what if we have two redundant but relevant features)

• therefore LR expected to outperform NB when given lots of training data
Naïve Bayes vs. logistic regression

non-asymptotic analysis [Ng & Jordan, *NIPS* 2001]

- consider convergence of parameter estimates; how many training instances are needed to get good estimates
  - naïve Bayes: $O(\log n)$
  - logistic regression: $O(n)$
  \[ n = \text{# features} \]

- naïve Bayes converges more quickly to its (perhaps less accurate) asymptotic estimates
- therefore NB expected to outperform LR with small training sets
Experimental comparison of NB and LR

Ng and Jordan compared learning curves for the two approaches on 15 data sets (some w/discrete features, some w/continuous features)
Experimental comparison of NB and LR

- logistic regression
- naïve Bayes

general trend supports theory
- NB has lower predictive error when training sets are small
- the error of LR approaches or is lower than NB when training sets are large
Discussion

• NB/LR is one case of a pair of generative/discriminative approaches for the same model class

• if modeling assumptions are valid (e.g. conditional independence of features in NB) the two will produce identical classifiers in the limit (# training instances → ∞)

• if modeling assumptions are not valid, the discriminative approach is likely to be more accurate for large training sets

• for small training sets, the generative approach is likely to be more accurate because parameters converge to their asymptotic values more quickly (in terms of training set size)

• Q: How can we tell whether our training set size is more appropriate for a generative or discriminative method?

A: Empirically compare the two.
THANK YOU

Some of the slides in these lectures have been adapted/borrowed from materials developed by Mark Craven, David Page, Jude Shavlik, Tom Mitchell, Nina Balcan, Elad Hazan, Tom Dietterich, and Pedro Domingos.