This lecture is based on (Ng & Jordan, 02) paper and some slides are based on Tom Mitchell’s slides.
Outline

- **Reminder:** Naive Bayes and Logistic Regression (MaxEnt)
- **Asymptotic Analysis**
  - What is better if you have an infinite dataset?
- **Non-asymptotic Analysis**
  - What is the rate of convergence of parameters?
  - More important: convergence of the expected error
- **Empirical evaluation**
- **Why this lecture?**
  - Nice and simple application of Large Deviation bounds we considered before
  - We will analyze specifically NB vs LogRegression, but “hope” it generalizes to other models (e.g., models for sequence labeling or parsing)
Discriminative vs Generative

- Training classifiers involves estimating $f: X \rightarrow Y$, or $P(Y|X)$

- **Discriminative classifiers** (conditional models)
  - Assume some functional form for $P(Y|X)$
  - Estimate parameters of $P(Y|X)$ directly from training data

- **Generative classifiers** (joint models)
  - Assume some functional form for $P(X|Y)$, $P(X)$
  - Estimate parameters of $P(X|Y)$, $P(X)$ directly from training data
  - Use Bayes rule to calculate $P(Y|X= x_i)$
Naive Bayes

- **Example:** assume $Y$ boolean, $X = \langle x_1, x_2, \ldots, x_n \rangle$, where $x_i$ are binary

- **Generative model:** Naive Bayes

  $$
  \hat{p}(x_i = 1|y = b) = \frac{s\{x_i = 1, y = b\} + l}{s\{y = b\} + 2l}
  $$

  $$
  \hat{p}(y = b) = \frac{s\{y = b\}}{\sum_j s\{y = j\}}
  $$

- **Classify new example $x$ based on ratio**

  $$
  \frac{\hat{p}(y = T|x)}{\hat{p}(y = F|x)} = \frac{\hat{p}(y = T)}{\hat{p}(y = F)} \prod_{i=1}^n \hat{p}(x_i|y = T) / \prod_{i=1}^n \hat{p}(x_i|y = F)
  $$

- **You can do it in log-scale**
Naive Bayes vs Logistic Regression

- **Generative model:** Naive Bayes

\[
\hat{p}(x_i = 1|y = b) = \frac{s\{x_i = 1, y = b\} + l}{s\{y = b\} + 2l}
\]

\[
\hat{p}(y = b) = \frac{s\{y = b\}}{\sum_j s\{y = j\}}
\]

- **Classify new example \( x \) based on ratio**

\[
\frac{\hat{p}(y = T|x)}{\hat{p}(y = F|x)} = \frac{\hat{p}(y = T) \prod_{i=1}^{n} \hat{p}(x_i|y = T)}{\hat{p}(y = F) \prod_{i=1}^{n} \hat{p}(x_i|y = F)}
\]

- **Logistic Regression:**

\[
\hat{p}(y = T|x; \beta, \theta) = \frac{1}{1 + \exp(-\sum_{i=1}^{n} \beta_i x_i - \theta)}
\]

- **Recall:** both classifiers are linear
What is the difference asymptotically?

Notation: let $\epsilon(h_{A,m})$ denote error of hypothesis learned via algorithm A, from $m$ examples

- If the Naive Bayes model is true: $\epsilon(h_{Dis,\infty}) = \epsilon(h_{Gen,\infty})$
- Otherwise $\epsilon(h_{Dis,\infty}) \leq \epsilon(h_{Gen,\infty})$
  - Logistic regression estimator is consistent:
    - $\epsilon(h_{Dis,m})$ converges to $\inf_{h \in \mathcal{H}} \epsilon(h)$
    - $\mathcal{H}$ is the class of all linear classifiers
    - Therefore, it is asymptotically better than the linear classifier selected by the NB algorithm
Proposition  Let $h_{\text{Dis}}$ be logistic regression in $n$-dimensions. Then with high probability

$$
\varepsilon(h_{\text{Dis}}) \leq \varepsilon(h_{\text{Dis},\infty}) + O\left(\sqrt{\frac{n}{m} \log \frac{m}{n}}\right)
$$

Thus, for $\varepsilon(h_{\text{Dis}}) < \varepsilon(h_{\text{Dis},\infty}) + \epsilon_0$ to hold with high probability (here, $\epsilon_0 > 0$ is some fixed constant), it suffices to pick $m = \Omega(n)$.

- Convergences to best linear classifier, in order of $n$ examples
  - follows from Vapnik’s structural risk bound (VC-dimension of $n$ dimensional linear separators is $n+1$)
Rate of convergence: Naive Bayes

- We will proceed in 2 stages:
  - Consider how fast parameters converge to their optimal values
    - *(we do not care about it, actually)*
  - We care: Derive how it corresponds to the convergence of the error to the asymptotical error

- The authors consider a continuous case (where input is continuous) but it is not very interesting for NLP
  - However, similar techniques apply
Lemma  Let any $\epsilon_1, \delta > 0$ and any $l \geq 0$ be fixed.
Assume that for some fixed $\rho_0 > 0$, we have that
$\rho_0 \leq p(y = T) \leq 1 - \rho_0$.
Let $m = O \left( \frac{1}{\epsilon_1^2} \log(n/\delta) \right)$.
Then with probability at least $1 - \delta$:
\[
\left| \hat{p}(x_i|y = b) - p(x_i|y = b) \right| \leq \epsilon_1 \\
\left| \hat{p}(y = b) - p(y = b) \right| \leq \epsilon_1
\]
for all $i = 1, \ldots, n$ and $b \in \mathcal{Y}$. 

Convergence of Parameters
Recall: Chernoff Bound

- Consider a binary random variable $X$ (e.g., the result of a coin toss) which has probability $p$ of being head (1).
- Consider $m$ samples, $\{x_1, x_2, \ldots\}$, drawn independently from $X$.
- The maximum likelihood estimate of $p$ is the relative frequency of $x_i = 1$. That is
  \[ \hat{p} = \frac{1}{m} \sum_{i} x_i. \]
- For all $p \in [0, 1]$, $\epsilon > 0$,
  \[ P[|p - \hat{p}| > \epsilon] \leq 2e^{-2m\epsilon^2}. \]
Recall: Union Bound

For any $n$ events \( \{A_1, A_2, \ldots A_n\} \) and for any distribution $P$ whose sample space includes all $A_i$,

\[
P[A_1 \cup A_2 \cup \ldots A_n] \leq \sum_i P[A_i]
\]
Proof of Lemma (no smoothing for simplicity)

- By the Chernoff’s bound, with probability at least:
  \[ 1 - \delta_1 = 1 - 2 \exp(-2\epsilon_1^2 m) \]
  the fraction of positive examples will be within \( \epsilon_1 \) of \( p(y = T) \):
  \[ |\hat{p}(y = b) - p(y = b)| \leq \epsilon_1 \]

- Therefore we have at least \( \gamma m \) positive and \( \gamma m \) negative examples
  \[ \gamma = \rho_0 - \epsilon_1 = \Omega(1) \quad (p(y = b) \text{ in } [\rho_0, 1- \rho_0]) \]

- By the Chernoff’s bound for every feature and class label (2n cases) with probability
  \[ \delta_2 = 2 \exp(-2\epsilon_1^2 \gamma m) \]
  \[ |\hat{p}(x_i|y = b) - p(x_i|y = b)| > \epsilon_1 \]

- We have one event with probability \( \delta_1 \) and \( 2n \) events with probabilities \( \delta_2 \), there joint probability is not greater than sum:
  \[ \delta_1 + 2n \delta_2 \leq \delta \]

- Solve this for \( m \), and you get
  \[ m = O \left( (1/\epsilon_1^2) \log(n/\delta) \right) \]
Implications

- With a number of samples logarithmic in $n$ (not linear as for the logistic regression!) the parameters of $h_{\text{Gen}}$ approach parameters of $h_{\text{Gen,} \infty}$.

- Are we done?
  - Not really: this does not automatically imply that the error $\varepsilon(h_{\text{Gen}})$ approaches $\varepsilon(h_{\text{Gen,} \infty})$ with the same rate.
Implications

- We need to show that $h_{\text{Gen}, \infty}$ and $h_{\text{Gen}}$ "often" agree if their parameters are close.
- We compare log-scores given by the models: $l_{\text{Gen}}(x)$ and $l_{\text{Gen}, \infty}(x)$.
- I.e.:

$$l_{\text{Gen}}(x) = \sum_{i=1}^{n} \log \frac{\hat{p}(x_i | y = T)}{\hat{p}(x_i | y = F)} + \log \frac{\hat{p}(y = T)}{\hat{p}(y = F)}$$
Theorem

Define

\[ G(\tau) = \Pr_{(x,y) \sim D}[(l_{Gen,\infty}(x) \in [0, \tau n] \land y = T) \lor (l_{Gen,\infty}(x) \in [-\tau n, 0] \land y = F)]. \]

Assume that for some fixed \( \rho_0 > 0 \), we have

\[ \rho_0 \leq p(y = T) \leq 1 - \rho_0, \]
\[ \rho_0 \leq p(x_i = 1 | y = b) \leq 1 - \rho_0 \text{ for all } i, b. \]

Then with high probability,

\[ \varepsilon(h_{Gen}) \leq \varepsilon(h_{Gen,\infty}) + G \left( O \left( \sqrt{\frac{1}{m} \log n} \right) \right). \]
Proof of Theorem (sketch)

- By the lemma (with high probability) the parameters of $h_{\text{Gen}}$ are within $O(\sqrt{(\log n)/m})$ of those of $h_{\text{Gen,} \infty}$

- It implies that every term in the sum $l_{\text{Gen}}(x)$ is also within $O(\sqrt{(\log n)/m})$ of the term in $l_{\text{Gen,} \infty}(x)$ and hence

$$|l_{\text{Gen}}(x) - l_{\text{Gen,} \infty}(x)| \leq O(n \sqrt{(\log n)/m})$$

- Let $\tau = O(\sqrt{(\log n)/m})$

- So $h_{\text{Gen}}$ and $h_{\text{Gen,} \infty}$ can have different predictions only if

  if $y = T$ and $l_{\text{Gen,} \infty}(x) \in [0, \tau n]$ ($l_{\text{Gen,} \infty}(x) \geq 0$, $l_{\text{Gen}}(x) \leq 0$)

  or if $y = F$ and $l_{\text{Gen,} \infty}(x) \in [-\tau n, 0]$

- Probability of this event is $G(\tau)$
Convergence of Classifiers

Theorem. Define
\[ G(\tau) = \Pr_{(x,y) \sim D}[(l_{Gen,\infty}(x) \in [0, \tau n] \land y = T) \lor (l_{Gen,\infty}(x) \in [-\tau n, 0] \land y = F)]. \]

Assume that for some fixed \( \rho_0 > 0 \), we have
\[
\begin{align*}
\rho_0 &\leq p(y = T) \leq 1 - \rho_0, \\
\rho_0 &\leq p(x_i = 1 | y = b) \leq 1 - \rho_0 \quad \text{for all } i, b
\end{align*}
\]

Then with high probability,
\[
\varepsilon(h_{Gen}) \leq \varepsilon(h_{Gen,\infty}) + G \left( O \left( \sqrt{\frac{1}{m} \log n} \right) \right).
\]

- \( G \) -- What is this fraction?
  - This is somewhat more difficult
Convergence of Classifiers

**Theorem**

Define

\[ G(\tau) = \Pr_{(x,y) \sim D}[(l_{\text{Gen},\infty}(x) \in [0, \tau n] \land y = T) \lor (l_{\text{Gen},\infty}(x) \in [-\tau n, 0] \land y = F)]. \]

Assume that for some fixed \( \rho_0 > 0 \), we have

\[ \rho_0 \leq p(y = T) \leq 1 - \rho_0, \]
\[ \rho_0 \leq p(x_i = 1 | y = b) \leq 1 - \rho_0 \] for all \( i, b \)

Then with high probability,

\[ \varepsilon(h_{\text{Gen}}) \leq \varepsilon(h_{\text{Gen},\infty}) + G \left( O \left( \sqrt{\frac{1}{m} \log n} \right) \right). \]

- **G** -- What is this fraction?
  - *This is somewhat more difficult*
What to do with this theorem

Proposition Suppose that, for at least an $\Omega(1)$ fraction of the features $i$ ($i = 1, \ldots, n$), it holds true that

$$|p(x_i = 1|y = T) - p(x_i = 1|y = F)| \geq \gamma$$

for some fixed $\gamma > 0$

Then

$$\mathbb{E}[l_{\text{Gen},\infty}(x)|y = T] = \Omega(n), \quad \text{and} \quad -\mathbb{E}[l_{\text{Gen},\infty}(x)|y = F] = \Omega(n).$$

- This is easy to prove, no proof but intuition:
  - A fraction of terms in $l_{\text{Gen},\infty}(x)$ have large expectation
  - Therefore, the sum has also large expectation
What to do with this theorem

Proposition Suppose that, for at least an $\Omega(1)$ fraction of the features $i$ ($i = 1, \ldots, n$), it holds true that

$$|p(x_i = 1|y = T) - p(x_i = 1|y = F)| \geq \gamma$$

for some fixed $\gamma > 0$

Then

$$E[l_{Gen,\infty}(x)|y = T] = \Omega(n), \quad \text{and} \quad -E[l_{Gen,\infty}(x)|y = F] = \Omega(n).$$

- But this is weaker than what we need:
  - We have that the expectation is “large”
  - We need that the probability of small values is low
- What about Chebyshev inequality?

$$\Pr[l_{Gen,\infty}(x) \leq E[l_{Gen,\infty}(x)] - t] \leq \frac{\text{Var}(l_{Gen,\infty}(x))}{t^2}$$

- They are not independent ... How to deal with it?
Corollary from the theorem

Corollary  Let the conditions of Theorem hold,

and suppose that \( G(\tau) \leq \epsilon_0/2 + F(\tau) \)

for some function \( F(\tau) \) (independent of \( n \))

that satisfies \( F(\tau) \to 0 \) as \( \tau \to 0 \), and some fixed \( \epsilon_0 > 0 \).

Then for \( \epsilon(h_{\text{Gen}}) \leq \epsilon(h_{\text{Gen,\infty}}) + \epsilon_0 \) to hold with high probability, it suffices to pick \( m = \Omega(\log n) \).

- Is this condition realistic?
  - Yes (e.g., we can show it for rather realistic conditions)
Empirical Evaluation (UCI dataset)

- Dashed line is logistic regression
- Solid line is Naives Bayes
Empirical Evaluation (UCI dataset)

- Dashed line is logistic regression
- Solid line is Naives Bayes
Empirical Evaluation (UCI dataset)

- Dashed line is logistic regression
- Solid line is Naives Bayes
Summary

- Logistic regression has lower asymptotic error
- ... But Naive Bayes needs less data to approach its asymptotic error
First Assignment

- I am still checking it, I will let you know by/on Friday
- Note though:
  - Do not perform multiple tests (model selection) on the final test set!
  - It is a form of cheating
This Friday I will distribute the first phase -- due after the Spring break

I will be away for the next 2 weeks:

- the first week (Mar, 9 – Mar, 15): I will be slow to respond to email
- I will be substituted for this week by:
  - Active Learning (Kevin Small)
  - Indirect Supervision (Alex Klementiev)
- Presentation by Ryan Cunningham on Friday

week Mar, 16 – Mar, 23 – no lectures:
- work on the project, send questions if needed