Outline

Theory:
- Probabilities:
  - Dependence, Independence, Conditional Independence
- Parameter estimation:
  - Maximum Likelihood Estimation (MLE)
  - Maximum a posteriori (MAP)
- Bayes rule
  - Naïve Bayes Classifier

Application:
Naive Bayes Classifier for
- Spam filtering
- “Mind reading” = fMRI data processing
Independence
Independence

Independent random variables:

\[ P(X, Y) = P(X)P(Y) \]
\[ P(X|Y) = P(X) \]

Y and X don’t contain information about each other. Observing Y doesn’t help predicting X. Observing X doesn’t help predicting Y.

Examples:

Independent: Winning on roulette this week and next week. Dependent: Russian roulette
Dependent / Independent

Independent X, Y

Dependent X, Y
Conditionally Independent

Conditionally independent:

\[ P(X, Y|Z) = P(X|Z)P(Y|Z) \]

Knowing Z makes X and Y independent

Examples:

Dependent: shoe size and reading skills
Conditionally independent: shoe size and reading skills given age

Storks deliver babies:

Highly statistically significant correlation exists between stork populations and human birth rates across Europe.
**London taxi drivers:** A survey has pointed out a positive and significant correlation between the number of accidents and wearing coats. They concluded that coats could hinder movements of drivers and be the cause of accidents. A new law was prepared to prohibit drivers from wearing coats when driving.

Finally another study pointed out that people wear coats when it rains…
Correlation ≠ Causation

I USED TO THINK CORRELATION IMPLIED CAUSATION.

THEN I TOOK A STATISTICS CLASS. NOW I DON’T.

SOUNDS LIKE THE CLASS HELPED.
Our first machine learning problem:

Parameter estimation: MLE, MAP

Estimating Probabilities
I have a coin, if I flip it, what’s the probability that it will fall with the head up?

Let us flip it a few times to estimate the probability:

The estimated probability is: $\frac{3}{5}$ “Frequency of heads”
The estimated probability is: \( \frac{3}{5} \) “Frequency of heads”

Questions:

(1) Why frequency of heads???
(2) How good is this estimation???
(3) Why is this a machine learning problem???

We are going to answer these questions
Why frequency of heads???

- Frequency of heads is exactly the maximum likelihood estimator for this problem.
- MLE has nice properties.
Maximum Likelihood Estimation
MLE for Bernoulli distribution

Data, $D =$

$D = \{X_i\}_{i=1}^{n}, \quad X_i \in \{H, T\}$

$P(\text{Heads}) = \theta, \quad P(\text{Tails}) = 1-\theta$

Flips are i.i.d.:
- Independent events
- Identically distributed according to Bernoulli distribution

MLE: Choose $\theta$ that maximizes the probability of observed data
MLE: Choose $\theta$ that maximizes the probability of observed data

$$\hat{\theta}_{MLE} = \arg \max_{\theta} P(D \mid \theta)$$

$$= \arg \max_{\theta} \prod_{i=1}^{n} P(X_i \mid \theta)$$

Independent draws

$$= \arg \max_{\theta} \prod_{i: X_i = H} \theta \prod_{i: X_i = T} (1 - \theta)$$

Identically distributed

$$= \arg \max_{\theta} \theta^{\alpha_H} (1 - \theta)^{\alpha_T}$$

$$J(\theta)$$
Maximum Likelihood Estimation

MLE: Choose $\theta$ that maximizes the probability of observed data

$$\hat{\theta}_{MLE} = \arg \max_\theta P(D \mid \theta)$$

$$= \arg \max_\theta \theta^{\alpha_H} (1 - \theta)^{\alpha_T}$$

$$J(\theta)$$

$$\frac{\partial J(\theta)}{\partial \theta} = \alpha_H \theta^{\alpha_H - 1} (1 - \theta)^{\alpha_T} - \alpha_T \theta^{\alpha_H} (1 - \theta)^{\alpha_T - 1} \bigg|_{\theta = \hat{\theta}_{MLE}} = 0$$

$$\alpha_H (1 - \theta) - \alpha_T \theta \bigg|_{\theta = \hat{\theta}_{MLE}} = 0$$

$$\hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

That's exactly the "Frequency of heads"
Question (2)

How good is this MLE estimation???

\[ \hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T} \]
How many flips do I need?

I flipped the coins 5 times: 3 heads, 2 tails

\[ \hat{\theta}_{MLE} = \frac{3}{5} \]

What if I flipped 26 heads and 24 tails?

\[ \hat{\theta}_{MLE} = \frac{26}{50} \]

• Which estimator should we trust more?
Simple bound

Let $\theta^*$ be the true parameter.

For $n = \alpha_H + \alpha_T$, and

$$\hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

For any $\varepsilon > 0$:

Hoeffding’s inequality:

$$P\left( | \hat{\theta} - \theta^* | \geq \varepsilon \right) \leq 2e^{-2n\varepsilon^2}$$
I want to know the coin parameter $\theta$, within $\epsilon = 0.1$ error with probability at least $1 - \delta = 0.95$.

How many flips do I need?

$$P(|\hat{\theta} - \theta^*| \geq \epsilon) \leq 2e^{-2n\epsilon^2} \leq \delta$$

Sample complexity:

$$n \geq \frac{\ln(2/\delta)}{2\epsilon^2}$$
Why is this a machine learning problem???

- improve their performance (accuracy of the predicted prob.)
- at some task (predicting the probability of heads)
- with experience (the more coins we flip the better we are)
What about continuous features?

Let us try Gaussians…

\[ p(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi \sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right) = \mathcal{N}_x(\mu, \sigma) \]
MLE for Gaussian mean and variance

Choose $\theta = (\mu, \sigma^2)$ that maximizes the probability of observed data

$$
\hat{\theta}_{MLE} = \arg \max_\theta P(D | \theta) \\
= \arg \max_\theta \prod_{i=1}^n P(X_i | \theta) \\
= \arg \max_\theta \prod_{i=1}^n \frac{1}{2\sigma^2} e^{-\frac{(X_i - \mu)^2}{2\sigma^2}} \\
= \arg \max_{\theta = (\mu, \sigma^2)} \frac{1}{2\sigma^2} e^{-\frac{\sum_{i=1}^n (X_i - \mu)^2}{2\sigma^2}} \\
= J(\theta)
$$

Independent draws

Identically distributed
MLE for Gaussian mean and variance

\[ \hat{\mu}_{MLE} = \frac{1}{n} \sum_{i=1}^{n} x_i \]

\[ \hat{\sigma}^2_{MLE} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 \]

**Note:** MLE for the variance of a Gaussian is **biased**
[Expected result of estimation is **not** the true parameter!]

Unbiased variance estimator:

\[ \hat{\sigma}^2_{unbiased} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 \]
What about prior knowledge? (MAP Estimation)
What about prior knowledge?

We know the coin is “close” to 50-50. What can we do now?

The Bayesian way...

Rather than estimating a single $\theta$, we obtain a distribution over possible values of $\theta$.
Prior distribution

What kind of prior distribution do we want to use?
- Represents expert knowledge (philosophical approach)
- Simple posterior form (engineer’s approach)

Uninformative priors:
- Uniform distribution

Conjugate priors:
- Closed-form representation of posterior
- $P(\theta)$ and $P(\theta|D)$ have the same form
Bayes Rule

Bayes rule is important for reverse conditioning.
Bayesian Learning

• Use Bayes rule:

\[ P(\theta \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \theta)P(\theta)}{P(\mathcal{D})} \]

• Or equivalently:

\[ P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta)P(\theta) \]

posterior likelihood prior
MLE vs. MAP

- **Maximum Likelihood estimation (MLE)**
  Choose value that maximizes the probability of observed data

\[
\hat{\theta}_{MLE} = \arg \max_\theta P(D|\theta)
\]

- **Maximum a posteriori (MAP) estimation**
  Choose value that is most probable given observed data and prior belief

\[
\hat{\theta}_{MAP} = \arg \max_\theta P(\theta|D)
\]
\[
= \arg \max_\theta P(D|\theta)P(\theta)
\]

When is MAP same as MLE?
MAP estimation for Binomial distribution

Coin flip problem: Likelihood is Binomial

\[ P(\mathcal{D} \mid \theta) = \binom{n}{\alpha_H} \theta^{\alpha_H} (1 - \theta)^{\alpha_T} \]

If the prior is Beta distribution,

\[ P(\theta) = \frac{\theta^{\beta_H - 1} (1 - \theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T) \]

⇒ posterior is Beta distribution

Beta function: \[ B(x, y) = \int_0^1 t^{x-1} (1 - t)^{y-1} \, dt \]
MAP estimation for Binomial distribution

Likelihood is Binomial: $P(D \mid \theta) = \binom{n}{\alpha_H} \theta^{\alpha_H} (1 - \theta)^{\alpha_T}$

Prior is Beta distribution: $P(\theta) = \frac{\theta^{\beta_H - 1} (1 - \theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T)$

$\Rightarrow$ posterior is Beta distribution

$$P(\theta \mid D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

P($\theta$) and P($\theta$|D) have the same form! [Conjugate prior]

$\hat{\theta}_{MAP} = \arg \max_{\theta} P(\theta \mid D) = \arg \max_{\theta} P(D \mid \theta)P(\theta)$

$$= \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2}$$
Beta distribution

More concentrated as values of $\alpha$, $\beta$ increase
As we get more samples, effect of prior is "washed out"
From Binomial to Multinomial

**Example**: Dice roll problem (6 outcomes instead of 2)

Likelihood is $\sim \text{Multinomial}(\theta = \{\theta_1, \theta_2, \ldots, \theta_k\})$

$$P(D \mid \theta) = \theta_1^{\alpha_1} \theta_2^{\alpha_2} \cdots \theta_k^{\alpha_k}$$

If prior is Dirichlet distribution,

$$P(\theta) = \frac{\prod_{i=1}^{k} \theta_i^{\beta_i - 1}}{B(\beta_1, \ldots, \beta_k)} \sim \text{Dirichlet}(\beta_1, \ldots, \beta_k)$$

Then posterior is Dirichlet distribution

$$P(\theta \mid D) \sim \text{Dirichlet}(\beta_1 + \alpha_1, \ldots, \beta_k + \alpha_k)$$

**For Multinomial, conjugate prior is Dirichlet distribution.**
Bayes Rule Application
Data
- Approximately 0.1% are infected
- Test detects all infections
- Test reports positive for 1% healthy people

Probability of having AIDS if test is positive:

\[
P(a = 1 | t = 1) = \frac{P(t = 1 | a = 1)P(a = 1)}{P(t = 1)}
\]

\[
= \frac{P(t = 1 | a = 1)P(a = 1)}{P(t = 1 | a = 1)P(a = 1) + P(t = 1 | a = 0)P(a = 0)}
\]

\[
= \frac{1 \cdot 0.001}{1 \cdot 0.001 + 0.01 \cdot 0.999} = 0.091
\]

Only 9%!
Improving the diagnosis

Use a follow-up test!
• Test 2 reports positive for 90% infections
• Test 2 reports positive for 5% healthy people

\[
P(a = 0|t_1 = 1, t_2 = 1) = \frac{P(t_1 = 1, t_2 = 1|a = 0)P(a = 0)}{P(t_1 = 1, t_2 = 1|a = 1)P(a = 1) + P(t_1 = 1, t_2 = 1|a = 0)P(a = 0)}
\]

\[
= \frac{0.01 \cdot 0.05 \cdot 0.999}{1 \cdot 0.9 \cdot 0.001 + 0.01 \cdot 0.05 \cdot 0.999} = 0.357
\]

\[
P(a = 1|t_1 = 1, t_2 = 1) = 0.643
\]

Why can’t we use Test 1 twice?
Outcomes are not independent but tests 1 and 2 are conditionally independent

\[
p(t_1, t_2|a) = p(t_1|a) \cdot p(t_2|a)
\]
The Naïve Bayes Classifier
Data for spam filtering

- date
- time
- recipient path
- IP number
- sender
- encoding
- many more features
Naïve Bayes Assumption

**Naïve Bayes assumption:** Features $X_1$ and $X_2$ are conditionally independent given the class label $Y$:

$$P(X_1, X_2|Y) = P(X_1|Y)P(X_2|Y)$$

More generally:

$$P(X_1...X_d|Y) = \prod_{i=1}^{d} P(X_i|Y)$$

How many parameters to estimate?

(X is composed of $d$ binary features, e.g. presence of word “earn” in a text. $Y$ has $K$ possible class labels)

$(2^d-1)K$ vs $(2-1)dK$
Naïve Bayes Classifier

Given:
- Class prior $P(Y)$
- $d$ conditionally independent features $X_1, ..., X_d$ given the class label $Y$
- For each $X_i$, we have the conditional likelihood $P(X_i | Y)$

Decision rule:
$$f_{NB}(x) = \arg \max_y P(x_1, \ldots, x_d | y) P(y)$$
$$= \arg \max_y \prod_{i=1}^{d} P(x_i | y) P(y)$$
Naïve Bayes Algorithm for discrete features

Training Data: \{ (X^{(j)}, Y^{(j)}) \}_{j=1}^n \quad X^{(j)} = (X_1^{(j)}, \ldots, X_d^{(j)})

\( n \times d \) dimensional features + class labels

\[^f_{NB}(x) = \arg\max_y \prod_{i=1}^d P(x_i|y)P(y) \quad \text{We need to estimate these probabilities!}\]

Estimate them with Relative Frequencies!

**For Class Prior**

\[ \hat{P}(y) = \frac{\# \{ j : Y^{(j)} = y \} }{n} \]

**For Likelihood**

\[ \frac{\hat{P}(x_i, y)}{\hat{P}(y)} = \frac{\# \{ j : X_i^{(j)} = x_i, Y^{(j)} = y \} / n}{\# \{ j : Y^{(j)} = y \} / n} \]

**NB Prediction for test data:**

\[ Y = \arg\max_y \hat{P}(y) \prod_{i=1}^d \frac{\hat{P}(x_i, y)}{\hat{P}(y)} \]
Subtlety: Insufficient training data

What if you never see a training instance where $X_1 = a$ when $Y = b$?

For example, there is no $X_1 = \text{‘Earn’}$ when $Y = \text{‘SpamEmail’}$ in our dataset.

\[
\Rightarrow P(X_1 = a, Y = b) = 0 \Rightarrow P(X_1 = a | Y = b) = 0
\]

\[
\Rightarrow P(X_1 = a, X_2...X_n | Y) = P(X_1 = a | Y) \prod_{i=2}^{d} P(X_i | Y) = 0
\]

Thus, no matter what the values $X_2,\ldots, X_d$ take:

\[
P(Y = b | X_1 = a, X_2,\ldots, X_d) = 0
\]

What now???
Case Study: Text Classification
Case Study: Text Classification

- Classify e-mails
  - $Y = \{\text{Spam, NotSpam}\}$
- Classify news articles
  - $Y = \{\text{what is the topic of the article?}\}$

What about the features $X$?
The text!
Article from rec.sport.hockey

Path: cantaloupe.srv.cs.cmu.edu!das-news.harvard.edu
From: xxx@yyy.zzz.edu (John Doe)
Subject: Re: This year’s biggest and worst (opinic
Date: 5 Apr 93 09:53:39 GMT

I can only comment on the Kings, but the most obvious candidate for pleasant surprise is Alex Zhitnik. He came highly touted as a defensive defenseman, but he’s clearly much more than that. Great skater and hard shot (though wish he were more accurate). In fact, he pretty much allowed the Kings to trade away that huge defensive liability Paul Coffey. Kelly Hrudey is only the biggest disappointment if you thought he was any good to begin with. But, at best, he’s only a mediocre goaltender. A better choice would be Tomas Sandstrom, though not through any fault of his own, but because some thugs in Toronto decided
**NB for Text Classification**

**P(X|Y) is huge!!!**
- Article at least 1000 words, \( X = \{X_1, \ldots, X_{1000}\} \)
- \( X_i \) represents \( i^{th} \) word in document, i.e., the domain of \( X_i \) is entire vocabulary, e.g., Webster Dictionary (or more).
  \( X_i \in \{1, \ldots, 50000\} \Rightarrow k50000^{1000} \) parameters....

**NB assumption helps a lot!!!**
- \( P(X_i=x_i|Y=y) \) is the probability of observing word \( x_i \) at the \( i^{th} \) position in a document on topic \( y \) \( \Rightarrow 1000k50000 \) parameters

\[
h_{NB}(x) = \arg \max_y P(y) \prod_{i=1}^{\text{LengthDoc}} P(x_i|y)
\]
Bag of words model

Typical additional assumption – **Position in document doesn’t matter**: $P(X_i = x_i | Y = y) = P(X_k = x_i | Y = y)$

- “Bag of words” model – order of words on the page ignored
- Sounds really silly, but often works very well! $\Rightarrow k50000$ parameters

$$\text{LengthDoc} \prod_{i=1}^{W} P(x_i | y) = \prod_{w=1}^{W} P(w | y)^{\text{count}_w}$$

When the lecture is over, remember to wake up the person sitting next to you in the lecture room.
Bag of words model

Typical additional assumption — **Position in document doesn’t matter**: \( P(X_i=x_i | Y=y) = P(X_k=x_i | Y=y) \)

- “Bag of words” model – order of words on the page ignored
- Sounds really silly, but often works very well!

\[
\prod_{i=1}^{\text{LengthDoc}} P(x_i | y) = \prod_{w=1}^{W} P(w | y)^{\text{count}_w}
\]
Bag of words approach

Our energy exploration, production, and distribution operations span the globe, with activities in more than 100 countries.

At TOTAL, we draw our greatest strength from our fast-growing oil and gas reserves. Our strategic emphasis on natural gas provides a strong position in a rapidly expanding market.

Our expanding refining and marketing operations in Asia and the Mediterranean Rim complement already solid positions in Europe, Africa, and the U.S.

Our growing specialty chemicals sector adds balance and profit to the core energy business.
Twenty news groups results

Given 1000 training documents from each group
Learn to classify new documents according to
which newsgroup it came from

comp.graphics  misc.forsale
comp.os.ms-windows.misc    rec.autos
comp.sys.ibm.pc.hardware  rec.motorcycles
comp.sys.mac.hardware  rec.sport.baseball
comp.windows.x         rec.sport.hockey

alt.atheism  sci.space
soc.religion.christian  sci.crypt
talk.religion.misc  sci.electronics
talk.politics.mideast  sci.med
talk.politics.misc
talk.politics.guns

Naïve Bayes: 89% accuracy
What if features are continuous?

Eg., character recognition: $X_i$ is intensity at $i^{th}$ pixel

**Gaussian Naïve Bayes (GNB):**

$$P(X_i = x \mid Y = y_k) = \frac{1}{\sigma_{ik}\sqrt{2\pi}} \frac{-(x-\mu_{ik})^2}{2\sigma_{ik}^2} e$$

Different mean and variance for each class $k$ and each pixel $i$.

Sometimes assume variance
- is independent of $Y$ (i.e., $\sigma_i$),
- or independent of $X_i$ (i.e., $\sigma_k$)
- or both (i.e., $\sigma$)
Example: GNB for classifying mental states

~1 mm resolution
~2 images per sec.
15,000 voxels/image
non-invasive, safe
measures Blood Oxygen Level Dependent (BOLD) response

[Mitchell et al.]
Learned Naïve Bayes Models – Means for $P(\text{BrainActivity} \mid \text{WordCategory})$

Pairwise classification accuracy: 78-99%, 12 participants

Tool words

Building words

[Mitchell et al.]
What you should know…

**Naïve Bayes classifier**
- What’s the assumption
- Why we use it
- How do we learn it
- Why is Bayesian (MAP) estimation important

**Text classification**
- Bag of words model

**Gaussian NB**
- Features are still conditionally independent
- Each feature has a Gaussian distribution given class
Further reading

Manuscript (book chapters 1 and 2)
http://alex.smola.org/teaching/berkeley2012/slides/chapter1_2.pdf

ML Books

Statistics 101
Thanks for your attention 😊
Many slides are taken from
• Tom Mitchel
  http://www.cs.cmu.edu/~tom/10701_sp11/slides
• Alex Smola
• Aarti Singh
• Eric Xing
• Xi Chen
• Wikipedia