MLE, MAP, Bayes Classification

Aykut Erdem
Dept. of Computer Engineering
Hacettepe University
This week

- Probabilities
  - Dependence, Independence, Conditional Independence
- Parameter estimation
  - Maximum Likelihood Estimation (MLE)
  - Maximum aposteriori (MAP)
- Bayes rule
  - Naïve Bayes Classifier
- Application
  - Spam filtering
  - “Mind reading” = fMRI data processing

Slides adapted from Alex Smola, Barnabás Póczos, Aarti Singh
Def: A sample space \( \Omega \) is the set of all possible outcomes of a (conceptual or physical) random experiment. (\( \Omega \) can be finite or infinite.)

Examples:
- \( \Omega \) may be the set of all possible outcomes of a dice roll \((1,2,3,4,5,6)\)
- Pages of a book opened randomly. \((1-157)\)
- Real numbers for temperature, location, time, etc
Recap: Events

We will ask the question:

What is the probability of a particular event?

Def: Event $A$ is a subset of the sample space $\Omega$

Examples:
What is the probability of
- the book is open at an odd number
- rolling a dice the number $<4$
- a random person’s height $X : a<X<b$
Def: Probability $P(A)$, the probability that event (subset) $A$ happens, is a function that maps the event $A$ onto the interval $[0, 1]$. $P(A)$ is also called the probability measure of $A$.

**Example:** What is the probability that the number on the dice is 2 or 4?
Kolmogorov Axioms

(i) Nonnegativity: $P(A) \geq 0$ for each $A$ event.

(ii) $P(\Omega) = 1$.

(iii) $\sigma$-additivity: For disjoint sets (events) $A_i$, we have

$$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

Consequences:

$$P(\emptyset) = 0.$$  
$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$  
$$P(A^c) = 1 - P(A).$$
Venn Diagram

\[ P(A \cup B) = P(A) + P(B) - P(A \cap B) \]
**Recap: Random Variables**

**Def:** Real valued **random variable** is a function of the outcome of a randomized experiment

\[ X : \Omega \rightarrow \mathbb{R} \]

\[
P(a < X < b) \equiv P(\omega : a < X(\omega) < b) \]
\[
P(X = a) \equiv P(\omega : X(\omega) = a) \]

**Examples:**

- **Discrete random variable examples (\(\Omega\) is discrete):**
  - \(X(\omega) = \) True if a randomly drawn person \((\omega)\) from our class \((\Omega)\) is female
  - \(X(\omega) = \) The hometown \(X(\omega)\) of a randomly drawn person \((\omega)\) from our class \((\Omega)\)
Recap: Discrete Distributions

• Bernoulli distribution: \( \text{Ber}(p) \)

\[
\Omega = \{ \text{head, tail} \} \quad X(\text{head}) = 1, \quad X(\text{tail}) = 0.
\]

\[
P(X = a) = P(\omega : X(\omega) = a) = \begin{cases} 
  p, & \text{for } a = 1 \\
  1 - p, & \text{for } a = 0
\end{cases}
\]

• Binomial distribution: \( \text{Bin}(n,p) \)

Suppose a coin with head prob. \( p \) is tossed \( n \) times. What is the probability of getting \( k \) heads and \( n-k \) tails?

\[
\Omega = \{ \text{possible } n \text{ long head/tail series} \}, \quad |\Omega| = 2^n
\]

\( K(\omega) = \text{number of heads in } \omega = (\omega_1, \ldots, \omega_n) \in \{\text{head, tail}\}^n = \Omega \)

\[
P(K = k) = P(\omega : K(\omega) = k) = \sum_{\omega : K(\omega) = k} p^k (1-p)^{n-k} = \binom{n}{k} p^k (1-p)^{n-k}
\]
Recap: Conditional Probability

$P(X|Y) = \frac{P(X,Y)}{P(Y)}$

$P(\text{flu|headache}) = \frac{P(\text{flu, headache})}{P(\text{headache})} = \frac{1/80}{1/80 + 7/80}$

<table>
<thead>
<tr>
<th></th>
<th>Flu</th>
<th>No Flu</th>
</tr>
</thead>
<tbody>
<tr>
<td>Headache</td>
<td>1/80</td>
<td>7/80</td>
</tr>
<tr>
<td>No Headache</td>
<td>1/80</td>
<td>71/80</td>
</tr>
</tbody>
</table>
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 of children and reading skills
Conditionally independent: shoe size of children and reading skills given age

Stork 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.
Conditional Independence

Formally: X is **conditionally independent** of Y given X

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

\[ P(\text{Accidents, Coats} | \text{Rain}) = P(\text{Accidents} | \text{Rain})P(\text{Coats} | \text{Rain}) \]

Equivalent to:

\[ (\forall x, y, z) P(X = x | Y = y, Z = z) = P(X = x | Z = z) \]

\[ P(\text{Thunder} | \text{Rain, Lightning}) = P(\text{Thunder} | \text{Lightning}) \]

**Note:** does NOT mean Thunder is independent of Rain

**But** given Lightning knowing Rain doesn’t give more info about Thunder
Conditional vs. Marginal Independence

- C calls A and B separately and tells them a number \( n \in \{1, \ldots, 10\} \)
- Due to noise in the phone, A and B each imperfectly (and independently) draw a conclusion about what the number was.
- A thinks the number was \( n_a \) and B thinks it was \( n_b \).
- Are \( n_a \) and \( n_b \) marginally independent?
  - No, we expect e.g. \( P(n_a = 1 | n_b = 1) > P(n_a = 1) \)
- Are \( n_a \) and \( n_b \) conditionally independent given \( n \)?
  - Yes, because if we know the true number, the outcomes \( n_a \) and \( n_b \) are purely determined by the noise in each phone.
  
  \[
P(n_a = 1 | n_b = 1, n = 2) = P(n_a = 1 | n = 2)
  \]
Parameter estimation: MLE, MAP
I have a coin, if I flip it, what’s the probability that it will fall with the head up?
Flipping a Coin

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:
Flipping a Coin

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:
Flipping a Coin

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”
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 \textit{maximum likelihood estimator} for this problem.
- MLE has nice properties (interpretation, statistical guarantees, simple)
Maximum Likelihood Estimation
MLE for Bernoulli distribution

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

$P(\text{Heads}) = \theta$, $P(\text{Tails}) = 1-\theta$
MLE for Bernoulli distribution

Data, $D =$

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

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

Flips are i.i.d.:
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 for Bernoulli distribution

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

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

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

$$\hat{\theta}_{MLE} = \arg\max_{\theta} P(D|\theta)$$
MLE: Choose $\theta$ that maximizes the probability of observed data

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

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

independent draws
Maximum Likelihood Estimation

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

\[ \hat{\theta}_{MLE} = \arg \max_{\theta} P(D|\theta) \]
\[ = \arg \max_{\theta} \prod_{i=1}^{n} P(X_i|\theta) \]
\[ = \arg \max_{\theta} \prod_{i:X_i=H} \theta \prod_{i:X_i=T} (1 - \theta) \]

independent draws

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

$$
\hat{\theta}_{MLE} = \arg \max_{\theta} P(D|\theta)
= \arg \max_{\theta} \prod_{i=1}^{n} P(X_i|\theta)
= \arg \max_{\theta} \prod_{i:X_i=H} \theta \prod_{i:X_i=T} (1 - \theta)
= \arg \max_{\theta} \theta^{\alpha_H} (1 - \theta)^{\alpha_T}
$$
MLE: Choose $\theta$ that maximizes the probability of observed data

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

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

independent draws

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

identically distributed

$$= \arg \max_{\theta} \theta^H (1 - \theta)^T$$

$J(\theta)$
Maximum Likelihood Estimation

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

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

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

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 30 heads and 20 tails?

$$\hat{\theta}_{MLE} = \frac{30}{50}$$

• Which estimator should we trust more?
• The more the merrier???
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:**

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

How many flips do I need?

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

Sample complexity:

$$n \geq \frac{\ln(2/\delta)}{2\varepsilon^2}$$
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...

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

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

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

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

= arg max $\prod_{i=1}^{n} \frac{1}{2\sigma^2} e^{-\frac{(X_i - \mu)^2}{2\sigma^2}}$

= arg max $\prod_{i=1}^{n} \frac{1}{2\sigma^2} e^{-\frac{\sum_{i=1}^{n} (X_i - \mu)^2}{2\sigma^2}}$

= arg max $\frac{1}{2\sigma^2} e^{-\frac{J(\theta)}{2\sigma^2}}$

$J(\theta)$
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 prior? What distribution do we want for a prior?
- 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
In order to proceed we will need:

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:

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

\[ \text{posterior} \propto \text{likelihood} \times \text{prior} \]

MAP estimation for Binomial distribution

Coin flip problem

Likelihood is Binomial

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

If the prior is Beta distribution,

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

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

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

\[ \hat{\theta}_{MAP} = \arg \max_{\theta} P(\theta \mid D) = \arg \max_{\theta} P(D \mid \theta)P(\theta) = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2} \]
More concentrated as values of $\alpha$, $\beta$ increase
Beta conjugate prior

\[ P(\theta) \sim \text{Beta}(\beta_H, \beta_T) \]
\[ P(\theta|D) \sim \text{Beta}(\beta_H + \alpha_H, \beta_T + \alpha_T) \]

As \( n = \alpha_H + \alpha_T \) increases

As we get more samples, effect of prior is “washed out”
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?
Example: Dice roll problem (6 outcomes instead of 2)

Likelihood is \(~\) 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.

Bayesians vs. Frequentists

You are no good when sample is small

You give a different answer for different priors
Recap: What about prior knowledge?
(MAP Estimation)
Recap: 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 $\theta$
Recap: Chain Rule & Bayes Rule

Chain rule:

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

Bayes rule:

\[ P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} \]
Recap: Bayesian Learning

D is the measured data
Our goal is to estimate parameter $\theta$

• 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
Recap: MAP estimation for Binomial distribution

In the coin flip problem:

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

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

then the posterior is Beta distribution
Recap: Beta conjugate prior

\[ P(\theta) \sim \text{Beta}(\beta_H, \beta_T) \quad P(\theta|D) \sim \text{Beta}(\beta_H + \alpha_H, \beta_T + \alpha_T) \]

As we get more samples, effect of prior is “washed out”
Application of Bayes Rule
AIDS test (Bayes rule)

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 weaker follow-up test!

- Approximately 0.1% are infected
- Test 2 reports positive for 90% infections
- Test 2 reports positive for 5% healthy people

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

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

\[
P(a = 1|t_1 = 1, t_2 = 1) = 0.643
\]
AIDS test (Bayes rule)

Why can’t we use Test 1 twice?

- Outcomes are not independent,
- but tests 1 and 2 conditionally independent (by assumption):

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

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

$$P(X_1...X_d|Y) = \prod_{i=1}^{d} P(X_i|Y)$$
Naïve Bayes Assumption, Example

Task: Predict whether or not a picnic spot is enjoyable

Training Data: \( X = (X_1, X_2, X_3, ..., X_d) \) \( \rightarrow \) \( Y \)

<table>
<thead>
<tr>
<th>Sky</th>
<th>Temp</th>
<th>Humid</th>
<th>Wind</th>
<th>Water</th>
<th>Forecast</th>
<th>EnjoySpt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunny</td>
<td>Warm</td>
<td>Normal</td>
<td>Strong</td>
<td>Warm</td>
<td>Same</td>
<td>Yes</td>
</tr>
<tr>
<td>Sunny</td>
<td>Warm</td>
<td>High</td>
<td>Strong</td>
<td>Warm</td>
<td>Same</td>
<td>Yes</td>
</tr>
<tr>
<td>Rainy</td>
<td>Cold</td>
<td>High</td>
<td>Strong</td>
<td>Warm</td>
<td>Change</td>
<td>No</td>
</tr>
<tr>
<td>Sunny</td>
<td>Warm</td>
<td>High</td>
<td>Strong</td>
<td>Cool</td>
<td>Change</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Naïve Bayes assumption: 
\[
P(X_1...X_d|Y) = \prod_{i=1}^{d} P(X_i|Y)
\]

How many parameters to estimate?

(X is composed of \( d \) binary features, \( Y \) 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 $X_1, \ldots, X_d$ given the class label $Y$
- For each $X_i$ feature, we have the conditional likelihood $P(X_i | Y)$

**Naïve Bayes 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

\[
\text{Training data: } \{(X^{(j)}, Y^{(j)})\}_{j=1}^{n}
\]

\[
X^{(j)} = (X_1^{(j)}, \ldots, X_d^{(j)})
\]

\[\text{n } d\text{-dimensional discrete features } + \text{ K class labels}\]

\[
f_{NB}(x) = \arg \max_y \prod_{i=1}^{d} P(x_i|y) P(y)\]

We need to estimate these probabilities!

Estimate them with MLE (Relative Frequencies)!
Naïve Bayes Algorithm for discrete features

\[ f_{NB}(x) = \arg \max_y \prod_{i=1}^{d} P(x_i | y) P(y) \]

We need to estimate these probabilities!

**Estimators**

For Class Prior

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

For Likelihood

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

**NB Prediction for Test Data:**

\[ X = (x_1, \ldots, x_d) \]

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

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

For example,

there is no $X_1$='Earn' when $Y$='SpamEmail' in our dataset.

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

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

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???
Naïve Bayes Alg — Discrete features

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

**Use your expert knowledge & apply prior distributions:**

- Add \( m \) “virtual” examples
- Same as assuming conjugate priors

**Assume priors:** \( Q(Y = b) \quad Q(X_i = a, Y = b) \)

**MAP Estimate:**

\[
\hat{P}(X_i = a | Y = b) = \frac{\{\#j : X^{(j)}_i = a, Y^{(j)} = b\} + mQ(X_i = a, Y = b)}{\{\#j : Y^{(j)} = b\} + mQ(Y = b)}
\]

\# virtual examples with \( Y = b \)
Case Study: Text Classification

- Classify emails
  - $Y = \{\text{Spam, NotSpam}\}$

- Classify news articles
  - $Y = \{\text{what is the topic of the article?}\}$

What are the features $X$?

The text!

Let $X_i$ represent $i^{th}$ word in the document
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
A problem: The support of $P(X|Y)$ is huge!

- Article at least 1000 words, $X=\{X_1,...,X_{1000}\}$
- $X_i$ represents $i^{th}$ word in document, i.e., the domain of $X_i$ is the entire vocabulary, e.g., Webster Dictionary (or more).

\[ X_i \in \{1,...,50000\} \Rightarrow K(1000^{50000} - 1) \]
parameters to estimate without the NB assumption....

\[
h_{MAP}(x) = \arg \max_{1 \leq k \leq K} P(Y = k) P(X_1 = x_1, \ldots, X_{1000} = x_{1000}|Y = k)\]
NB for Text Classification

\[ X_i \in \{1, \ldots, 50000\} \Rightarrow K(1000^{50000} - 1) \text{ parameters to estimate.} \]

**NB assumption helps a lot!!!**

If \( P(X_i=x_i | Y=y) \) is the probability of observing word \( x_i \) at the \( i^{th} \) position in a document on topic \( y \)

\[ \Rightarrow 1000K(50000-1) \text{ parameters to estimate with NB assumption} \]

**NB assumption helps, but still lots of parameters to estimate.**

\[ h_{NB}(x) = \arg \max_y P(y) \prod_{i=1}^{\text{LengthDoc}} P(X_i = x_i | y) \]
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
  - The document is just a bag of words: i.i.d. words
  - Sounds really silly, but often works very well!

⇒ \( K(50000-1) \) parameters to estimate

The probability of a document with words \( x_1, x_2, \ldots \)

\[
\prod_{i=1}^{LengthDoc} P(x_i | y) = \prod_{w=1}^{W} P(w | y)^{count_w}
\]
When the lecture is over, remember to wake up the person sitting next to you in the lecture room.
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?

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

**Gaussian Naïve Bayes (GNB):**

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

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., $\sigma$)
Estimating parameters: Y discrete, X_i continuous

\[
h_{NB}(x) = \arg \max_y P(y) \prod_i P(X_i = x_i | y)
\]

\[
\approx \arg \max_k \hat{P}(Y = k) \prod_i \mathcal{N}(\hat{\mu}_{ik}, \hat{\sigma}_{ik})
\]

\[
\hat{\mu}_{MLE} = \frac{1}{N} \sum_{j=1}^{N} x_j
\]

\[
\hat{\sigma}_{unbiased}^2 = \frac{1}{N - 1} \sum_{j=1}^{N} (x_j - \hat{\mu})^2
\]
Estimating parameters: Y discrete, X\textsubscript{i} continuous

Maximum likelihood estimates:

\[ \hat{\mu}_{MLE} = \frac{1}{N} \sum_{j=1}^{N} x_j \]

\[ \hat{\mu}_{ik} = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j X^j_i \delta(Y^j = y_k) \]

\[ \hat{\sigma}_{unbiased}^2 = \frac{1}{N-1} \sum_{j=1}^{N} (x_j - \hat{\mu})^2 \]

\[ \hat{\sigma}_{ik}^2 = \frac{1}{\sum_j \delta(Y^j = y_k) - 1} \sum_j (X^j_i - \hat{\mu}_{ik})^2 \delta(Y^j = y_k) \]
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

[Mitchell et al.]
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