CS 598 Machine Learning for Signal Processing

Probability, Statistics & Parameter Estimation

28 August 2015
Logistics

• Did everyone get the class email?
  • If not, send me your NetID so that I can add you to the mailing list

• Is there a waiting list to register for the class?
  • Sorry no, just keep trying to register

• Class recordings are available for registered students at:
  • https://recordings.engineering.illinois.edu:8443/ess/portal/section/242d0f51-7fa8-49d2-aa4c-b2b78701dc10
    • Remember attendance counts!
Today’s refresher

- Probability
- Statistics
- Parameter Estimation
Probability

- Probit
  - Measure of legal authority/nobility
    - Passed muster in the middle ages

- Probability
  - Measure of belief/likelihood
    - Passes muster today
Goals of probability

• Characterize stochastic processes
  • How do dice roll?
  • What am I more likely to say next?

• Indicate belief given evidence
  • The suspect was nearby and there are feathers on his clothes. Was he the chicken thief?
An example

- We start picking oranges, apples and bananas, from the two boxes below
  - Pick 40% from red box, 60% from green box
The random variables

- The box: $B = \{r, g\}$
- The fruit: $F = \{a, o, b\}$
  - What are their probabilities?
Box probabilities

- Obviously:
  - $P(B == g) = 6/10$
  - $P(B == r) = 4/10$
  - $P(\cdot) \in [0,1]$
Asking questions

- What is the probability of picking an apple?
- If we pick an orange, what is the probability that it came out of the green box?
Keeping track

- Keep track of $N$ experiments in a table
- $N$ is large, even infinite

\[
\begin{array}{|c|c|c|c|}
\hline
& \text{Apple} & \text{Banana} & \text{Orange} \\
\hline
\text{Green Box} & n_{ga} & n_{gb} & n_{go} \\
\hline
\text{Red Box} & n_{ra} & n_{rb} & n_{ro} \\
\hline
\text{Any box} & n_a & n_b & n_o \\
\hline
\end{array}
\]
### Single variable probabilities

\[
P(B \equiv i) = \frac{n_i}{N} \\
P(F \equiv j) = \frac{n_j}{N}
\]

\[
B
\]

**Green Box**

- Apple: \( n_{ga} \)
- Banana: \( n_{gb} \)
- Orange: \( n_{go} \)
- Any fruit: \( n_g \)

**Red Box**

- Apple: \( n_{ra} \)
- Banana: \( n_{rb} \)
- Orange: \( n_{ro} \)
- Any fruit: \( n_r \)

**Any box**

- Apple: \( n_a \)
- Banana: \( n_b \)
- Orange: \( n_o \)
Joint probabilities

\[ P(B = i, F = j) = \frac{n_{ij}}{N} \]

\[ P(B = i, F = j) = P(F = j, B = i) \]

<table>
<thead>
<tr>
<th></th>
<th>Apple</th>
<th>Banana</th>
<th>Orange</th>
<th>Any fruit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Green Box: B</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n_{ga}</td>
<td>n_{gb}</td>
<td>n_{go}</td>
<td>( n_g )</td>
<td></td>
</tr>
<tr>
<td>Red Box: F</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n_{ra}</td>
<td>n_{rb}</td>
<td>n_{ro}</td>
<td>( n_r )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>n_{a}</td>
<td>n_{b}</td>
<td>n_{o}</td>
<td>( n )</td>
</tr>
</tbody>
</table>
The sum rule

\[ n_i / N = \left( n_{ia} + n_{ib} + n_{io} \right) / N \]

\[ P(B == i) = \sum_{\forall j} P(B == i, F == j) \]

<table>
<thead>
<tr>
<th>Fruit</th>
<th>Green Box</th>
<th>Red Box</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apple</td>
<td>( n_{ga} )</td>
<td>( n_{ra} )</td>
</tr>
<tr>
<td>Banana</td>
<td>( n_{gb} )</td>
<td>( n_{rb} )</td>
</tr>
<tr>
<td>Orange</td>
<td>( n_{go} )</td>
<td>( n_{ro} )</td>
</tr>
<tr>
<td>Any fruit</td>
<td>( n_{g} )</td>
<td>( n_{r} )</td>
</tr>
<tr>
<td></td>
<td>( n_{a} )</td>
<td>( n_{b} )</td>
</tr>
</tbody>
</table>
Conditional probability

\[ P(F == j \mid B == i) = \frac{n_{ij}}{n_i} \]

<table>
<thead>
<tr>
<th></th>
<th>Apple</th>
<th>Banana</th>
<th>Orange</th>
<th>Any fruit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Green Box</td>
<td>(n_{ga})</td>
<td>(n_{gb})</td>
<td>(n_{go})</td>
<td>(n_g)</td>
</tr>
<tr>
<td>Red Box</td>
<td>(n_{ra})</td>
<td>(n_{rb})</td>
<td>(n_{ro})</td>
<td>(n_r)</td>
</tr>
<tr>
<td></td>
<td>(n_a)</td>
<td>(n_b)</td>
<td>(n_o)</td>
<td></td>
</tr>
</tbody>
</table>
The product rule

\[ P(B = i, F = j) = \frac{n_{ij}}{N} = \frac{n_{ij}}{n_i} \frac{n_i}{N} = P(F = j | B = i) P(B = i) \]

\[ F \]

<table>
<thead>
<tr>
<th></th>
<th>Apple</th>
<th>Banana</th>
<th>Orange</th>
<th>Any fruit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Green Box</td>
<td>( n_{ga} )</td>
<td>( n_{gb} )</td>
<td>( n_{go} )</td>
<td>( n_g )</td>
</tr>
<tr>
<td>Red Box</td>
<td>( n_{ra} )</td>
<td>( n_{rb} )</td>
<td>( n_{ro} )</td>
<td>( n_r )</td>
</tr>
<tr>
<td></td>
<td>( n_a )</td>
<td>( n_b )</td>
<td>( n_o )</td>
<td></td>
</tr>
</tbody>
</table>
The two basic rules

- **Sum Rule:**

\[ P(X) = \sum_{Y} P(X,Y) \]

- **Product Rule:**

\[ P(X,Y) = P(Y \mid X)P(X) \]
Bayes theorem

- From product rule & joint symmetry

\[
P(Y | X) = \frac{P(X | Y)P(Y)}{P(X)}
\]

\( P(Y | X) \) \( \text{Posterior} \)
\( P(X | Y)P(Y) \) \( \text{Likelihood} \)
\( P(X) \) \( \text{Prior} \)
\( \text{Normalizing constant} \)

- Will answer most of your questions!
Independence

- If:

\[ P(B = i, F = j) = P(B = i)P(F = j) \]

- Then \( B \) and \( F \) are independent

- Also means, via the product rule, that:

\[ P(F | B) = P(F) \]

- If both boxes had the same fraction of fruits, then we would have independence
Back to the fruit

- What’s the probability of picking a banana?
  - Sum rule: \( P(b) = P(b, r) + P(b, g) \)
Back to the fruit

• What’s the probability of the red box given that I picked an apple?
  • Bayes rule: $P(r \mid a) = \frac{P(a \mid r)P(r)}{P(a)}$
Schools of thought

• Frequentists
  • Probabilities are interpretations of frequencies of occurrence in experiments
    • There can only be one solution!

• Bayesians
  • Probabilities are a degree of belief, not a result of a counting experiment
    • What’s the distribution of the parameter? The priors?
Why belief?

• “Will a meteor hit earth?”
  • Frequentist: Let us wait until $N$ is large ...

• Using a Bayesian treatment we can find a likelihood given the evidence, not the data
  • But that requires models, priors, assumptions, ... More later
A practical application

Statistics in the Real World: The Search for the USS Scorpion

http://www.youtube.com/watch?v=U9-G-noZrwc
Getting lost? Don’t worry

• Probability is super tricky
  • Even seasoned professionals get it wrong!
    • E.g. the Monty Hall problem

http://marilynvossavant.com/game-show-problem/
## Quick answer

<table>
<thead>
<tr>
<th>Pick door 1 and switch</th>
<th>Door 1</th>
<th>Door 2</th>
<th>Door 3</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st case</td>
<td>Car</td>
<td>Goat</td>
<td>Goat</td>
<td>Switch &amp; lose</td>
</tr>
<tr>
<td>2nd case</td>
<td>Goat</td>
<td>Car</td>
<td>Goat</td>
<td>Switch &amp; win</td>
</tr>
<tr>
<td>3rd case</td>
<td>Goat</td>
<td>Goat</td>
<td>Car</td>
<td>Switch &amp; win</td>
</tr>
<tr>
<td>4th case</td>
<td>Car</td>
<td>Goat</td>
<td>Goat</td>
<td>Switch &amp; win</td>
</tr>
<tr>
<td>5th case</td>
<td>Goat</td>
<td>Car</td>
<td>Goat</td>
<td>Switch &amp; lose</td>
</tr>
<tr>
<td>6th case</td>
<td>Goat</td>
<td>Goat</td>
<td>Car</td>
<td>Switch &amp; lose</td>
</tr>
</tbody>
</table>

Pick door 1 and stay

<table>
<thead>
<tr>
<th>Door 1</th>
<th>Door 2</th>
<th>Door 3</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Car</td>
<td>Goat</td>
<td>Goat</td>
<td>Switch &amp; win</td>
</tr>
<tr>
<td>Goat</td>
<td>Car</td>
<td>Goat</td>
<td>Switch &amp; win</td>
</tr>
<tr>
<td>Goat</td>
<td>Goat</td>
<td>Car</td>
<td>Switch &amp; lose</td>
</tr>
<tr>
<td>Goat</td>
<td>Goat</td>
<td>Car</td>
<td>Switch &amp; lose</td>
</tr>
</tbody>
</table>
Continuous distributions

- What if we have infinite colors of boxes, and infinite types of fruit?
Same(ish) rules (harder proofs)

- **Sum rule:** \( P(x) = \int P(x, y) dy \)

- **Product rule:** \( P(x, y) = P(y | x)P(x) \)

- **Bayes rule:** \( P(x | y) = \frac{P(y | x)P(x)}{P(y)} \)
Some properties

- Integration to unity

\[ \int_{-\infty}^{\infty} P(x) = 1 \]

  - You’ll be amazed how many get this wrong!

- Probabilities are real and non-negative

\[ P(x) \in \mathbb{R} \quad P(x) \geq 0 \]

  - Well, they don’t have to be. More on that later ...
Useful operations

• **Expectation:** \( \mathbb{E}(f(x)) = \int P(x)f(x)\,dx \)

• **Conditional expectation:** \( \mathbb{E}_x(f(x) \mid y) = \int P(x \mid y)f(x)\,dx \)

• **Variance:** \( \text{var}(f(x)) = \mathbb{E}\left[\left(f(x) - \mathbb{E}[f(x)]\right)^2\right] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2 \)

• **Covariance:** \( \text{cov}[x, y] = \mathbb{E}_{x,y}(xy) - \mathbb{E}(x)\mathbb{E}(y) \)
Popular distributions

- We’ll be seeing a lot of:
  - The Gaussian
    - Used pretty much everywhere
  - The Laplacian
    - Used for sparse models
  - The Dirichlet
    - Used for compositional models
  - The Exponential Family
    - Very useful properties!
The Gaussian

- Also known as the Normal distribution or the bell curve

\[ \mathcal{N}(\mathbf{x}; \mu, \Sigma) = \frac{1}{\sqrt{2\pi^D |\Sigma|}} e^{-\frac{1}{2}(\mathbf{x} - \mu)^\top \Sigma^{-1}(\mathbf{x} - \mu)} \]

\( \mathbf{x} \in \mathbb{R}^D \)

One-dimensional Gaussians

Two-dimensional Gaussians
Why the Gaussian?

- Makes the Euclidean distance a distribution

\[ \mathcal{N}(x; \mu, \sigma) = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]

- If you assume squared Euclidean errors, then you are using a Gaussian
The Gaussian parameters

\[ \mathcal{N}(x; \mu, \Sigma) = \frac{1}{\sqrt{2\pi^D |\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)} \quad x \in \mathbb{R}^D \]

- The mean: \( E(x) = \mu \)
- The covariance: \( \text{cov}(x) = \Sigma \)
- The mode: \( \text{mode}(x) = \mu \)
Special case

Fig 1.0 The Extended Bell Curve.

– by Tang Yau Hoong
The Laplacian

- Sharper than the Gaussian
- Uses absolute distance, not Euclidean

\[ P(x; \mu, b) = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}} \]

- Mean: \( \mu \)
- Variance: \( 2b^2 \)
- Mode: \( \mu \)
Beta/Dirichlet distributions

- Defined on a simplex
  - $x_1 + x_2 + x_3 + ... = 1$
  
  $$P(x; a) = \frac{\prod \Gamma(a_i)}{\Gamma(\sum a_i)} \prod x_i^{a_i-1}$$

- For 1D the Dirichlet is the Beta

- Mean: $E[x_i] = a_i / a_0$

- Variance: $\text{cov}[x_i, x_j] = -a_i a_j / a_0^2(a_0 + 1)$

- Mode: $x_i = (a_i - 1) / (a_0 - K)$
The exponential family

- Any distribution that can be written as:
  \[ P(x; \eta) = h(x) g(\eta) e^{\eta^\top u(x)} \]

- \( \eta \) contains the natural parameters
- \( u(x) \) is some function of \( x \)
- \( g(\eta) \) is just for normalization
Gaussian example

\[ P(x; \eta) = h(x) g(\eta) e^{\eta^\top u(x)} \]

\[ u(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix}, \quad h(x) = (2\pi)^{-1/2} \]

\[ \eta = \begin{bmatrix} \mu / \sigma^2 \\ -1 / 2\sigma^2 \end{bmatrix}, \quad g(\eta) = (-2\eta_2)^{1/2} e^{\eta_1^2 / 4\eta_2} \]

\[ P(x; h) = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{1}{2\sigma^2} x^2 + \frac{\mu}{\sigma^2} x - \frac{1}{2\sigma^2} \mu^2} = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]
Why this mess???

- Allow us to see a broader picture

- Exponential distributions have convenient properties
  - Sufficiency
    - You won’t need more parameters for more data
  - Conjugate priors
    - Make life easy when we perform parameter estimation (more later)
Information theory

• Entropy

\[ H(x) = -\int P(x) \log P(x) \, dx \quad \text{or} \quad -\sum_x P(x) \log P(x) \]

\[ H(x, y) = -\int \int P(x, y) \log P(x, y) \, dx \, dy \quad \text{or} \quad -\sum_x \sum_y P(x, y) \log P(x, y) \]

• A measure of information in a distribution

A fair die, \( H = 1.79 \)
There is a lot of uncertainty therefore more information

A heavily biased die, \( H = 0 \)
no message to convey
Information theory

- Mutual information
  - Measures amount of shared information
    \[ I(x, y) = H(x) + H(y) - H(x, y) \]
  - If 0 then \( x, y \) are independent

- Kullback-Leibler divergence
  - A pseudo-distance for distributions
    \[ D(p \mid q) = \sum p_i \log \frac{p_i}{q_i} \quad \text{or} \quad \int p(x) \log \frac{p(x)}{q(x)} \, dx \]
    \[ D(P(x, y) \mid \mid P(x)P(y)) = I(x, y) \]
  - If 0 then \( p \) and \( q \) are the same
Entropy types

- \( H(X) \)
- \( H(Y) \)
- \( H(X|Y) \)
- \( I(X;Y) \)
- \( H(Y|X) \)
- \( H(X,Y) \)
Parameter estimation

• So what do we do with distributions?
  • We like to explain data with them

• To do so we need parameter estimation
  • Find the distribution parameters that result in explaining the observed data best
  • Various ways to go about it
Parameter estimation

• Given some independent samples:

\[ X = \{x_1, x_2, \ldots, x_N \} \]

• and a model:

\[ P(X; \theta) \]

• Find the parameters \( \theta \)
Maximum likelihood

• The overall likelihood is:

\[ P(X; \theta) = P(x_1, x_2, \ldots, x_N; \theta) = \prod_i P(x_i; \theta) \]

• We want to find:

\[ \theta_{ML} = \arg \max_{\theta} \prod_i P(x_i; \theta) \]

• We can use straightforward solving
Maximum likelihood

- Set the derivative to zero:

\[
\frac{\partial \prod_i P(x_i; \theta)}{\partial \theta} = 0
\]

- Go to the log domain to remove product:

\[
\frac{\partial \log \prod_i P(x_i; \theta)}{\partial \theta} = \sum_i \frac{\partial \log P(x_i; \theta)}{\partial \theta} = \sum_i \frac{1}{P(x_i; \theta)} \frac{\partial P(x_i; \theta)}{\partial \theta} = 0
\]

- Substitute your \( P \) and solve
Example

• Mean of Gaussian distributed data
  • Define the model:
    \[ P(\mathbf{x}; \mu, \sigma^2) = \prod_{i=1}^{N} \mathcal{N}(\mathbf{x}; \mu, \sigma^2) \]
  • Form log-likelihood:
    \[ \log P(\mathbf{x}; \mu, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{i=1}^{N} (x_n - \mu)^2 - \frac{N}{2} \log \sigma^2 - \frac{N}{2} \log 2\pi \]
  • Set derivative to zero and solve:
    \[ \frac{\partial \log P(\mathbf{x}; \mu, \sigma^2)}{\partial \mu} = -\frac{1}{2\sigma^2} \sum_{i=1}^{N} \frac{\partial (x_n - \mu)^2}{\partial \mu} = 0 \Rightarrow \mu = \frac{1}{N} \sum_{i=1}^{N} x_i \]
Wait a minute!

• All that to prove the obvious?

• Yes, it is tedious
  • In many cases the answer will be obvious
    • But keep in mind that looks might be deceiving!

• In other cases the answer will not be easy
  • Requiring numerical/approximate optimization
A couple of ML properties

- The ML estimate is (usually) asymptotically Gaussian distributed and:

\[
\lim_{N \to \infty} E[\theta_{ML}] = \theta_{true} \quad \text{and} \quad \lim_{N \to \infty} E\left[\left\| \theta_{ML} - \theta_{true} \right\|^2 \right] = 0
\]
Maximum a posteriori (MAP)

- Sometimes we have a prior belief
  - E.g. we believe the answer should be close to a value
  - Maximum likelihood doesn’t incorporate that
  - MAP does

- Same setup as before but in addition to $P(x;\theta)$ we also have a $P(\theta)$
MAP estimation

- We use Bayes’ theorem and we now maximize:

\[ P(\theta | x) = \frac{P(\theta)P(x | \theta)}{P(x)} \]

- The denominator is constant so we only have to maximize the numerator:

\[ \theta_{MAP} = \arg \max_{\theta} P(\theta)P(x | \theta) \]

- Same story as before ...
MAP estimation example

- Estimate the mean, but use a prior:

\[
P(x; \mu, \sigma^2) = \prod_{i=1}^{N} \mathcal{N}(x; \mu, \sigma^2), \quad P(\mu; \mu_0, \sigma^2_\mu) = \mathcal{N}(\mu, \mu_0, \sigma^2_\mu)
\]

- Take log, differentiate, solve:

\[
\frac{\partial}{\partial \mu} \log \prod_{i=1}^{N} P(x_i | \mu) P(\mu) = 0
\]

\[
\sum_{i=1}^{N} \frac{1}{\sigma^2} (x_i - \mu) - \frac{1}{\sigma^2_\mu} (\mu - \mu_0) = 0
\]

\[
\Rightarrow \mu_{MAP} = \frac{\mu_0 + \frac{\sigma^2}{\sigma^2_\mu} \sum_{i=1}^{N} x_i}{\frac{\sigma^2}{\sigma^2_\mu} + 1}
\]
MAP vs. ML

- If $P(\theta)$ is uniform then MAP $==$ ML
- Otherwise they will most likely not coincide
Bayesian inference

- Bayesian inference doesn’t care about the optimal value, it cares about it’s distribution.
Example estimation

- Same setup as in the MAP case:

\[ P(x; \mu, \sigma^2) = \prod_{i=1}^{N} N(x; \mu, \sigma^2), \quad P(\mu; \mu_0, \sigma^2_\mu) = N(\mu, \mu_0, \sigma^2_\mu) \]

- We now find the distribution of the mean:

\[ P(\mu | X) = \frac{P(X | \mu)P(\mu)}{P(X)} = \ldots = N(\mu, \mu_N, \sigma^2_N) \]

\[ \mu_N = \frac{N\sigma^2_0 \mathbb{E}[x] + \sigma^2 \mu_0}{N\sigma^2_0 + \sigma^2}, \quad \sigma^2_N = \frac{\sigma^2 \sigma^2_0}{N\sigma^2_0 + \sigma^2} \]

- Which is also Gaussian!
Obtaining the estimate

- For different values of $N$ we obtain a different distribution of the parameter we estimate.
  - The bigger the $N$ the more sharp the distribution.
And that was a clean case

- Often the distributions don’t work out
- We resort to numerical solutions
  - Usually sampling (Monte Carlo, etc.)
Other methods

• Maximum entropy estimation
  • Choose model that maximizes entropy
    • Least committal approach

• Expectation-Maximization
  • Useful for mixture models
  • We’ll cover in detail later
Recap

- **Probability**
  - sum/product/Bayes rules

- **Distributions**
  - Gaussian, Laplacian, Dirichlet

- **Information theory**
  - Entropy, Mutual Info, KL divergence

- **Parameter estimation**
  - ML, MAP, Bayesian
Too much information?

• You are not supposed to master all this
  • We will be encountering these ideas later
  • This lecture should serve as a reference
Some more reading

- Get textbook from class page
  - UIUC network access only

- Probability basics
  - Appendix 1 of textbook

- Parameter estimation
  - Section 2.5 of textbook
Next week

• Signals refresher
  • “All of DSP in a lecture”