Some references for probability

(We will be closest to chapter two of K&F—posted)

Wasserman, "All of statistics"
(Vision lab has a few copies that we can lend out).

Forsyth and Ponce chapter
http://luthuli.cs.uiuc.edu/~daf/book/bookpages/pdf/probability.pdf

Your favorite intro to probability book (e.g., "Mathematical statistics and Data Analysis," by John Rice.)

Google (and WikiPedia)

Probability review

Formulas that you should be very comfortable with are marked by *.

Interpretations of probability

1) Representation of expected frequency

2) Degree of belief

Basic terminology and rules

Space of outcomes (often denoted by Ω)

Event (subset of Ω)

Denote the space of measurable events (one we want to assign a probability to) by $\mathcal{S}$.

$\mathcal{S}$ must include $\varnothing$ and Ω

$\mathcal{S}$ is closed under set operations

$\alpha,\beta \in \mathsf{S} \Rightarrow \alpha \cup \beta \in \mathsf{S}$, $\alpha \cap \beta \in \mathsf{S}$, $\alpha^C = \Omega - \alpha \in \mathsf{S}$, etc.

Basic terminology and rules

A probability distribution P over (Ω,$\mathcal{S}$) is a mapping from events in S to real values such that

If $a \in S$, $P(a) \geq 0$

$P(\Omega) = 1$

If $\alpha, \beta \in S$, and a$\cap$b=$\varnothing$, then $P(a \cup b) = P(a) + P(b)$

Basic terminology and rules

We can further derive additional familiar facts

$$\text{If } a \in S, \quad P(a) \in [0,1]$$

$$P(\varnothing) = 0$$

$$\text{If } \alpha, \beta \in S, \text{ then } P(a \cup b) = P(a) + P(b) - P(a \cap b)$$

The probabilities over disjoint sets that cover P(Ω) sum to 1.

Basic terminology and rules

Conditional probability (definition)

$$P\left(A \middle| B\right) \equiv \frac{P(A \cap B)}{P(B)}$$ *

Example, what is the probability that you have rolled 2, given that you know you have rolled a prime number?

Basic terminology and rules

Conditional probability (definition)

$$P\left(A \middle| B\right) \equiv \frac{P(A \cap B)}{P(B)}$$

Applying a bit of algebra,

$$P\left(A \cap B\right) = P(A)P(B|A)$$

In general, we have the chain (product) rule

$$P\left(A_1 \cap A_2\right) = P(A_1)P(A_2|A_1)$$ *

$$P\left(A_1 \cap A_2 \cap \;\; .... \;\; A_N\right) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \; .... \; P(A_N|A_1 \cap A_2 \cap \;\; .... \;\; A_{N-1})$$

Basic terminology and rules

From before, we define conditional probability by

$$P\left(A \middle| B\right) \equiv \frac{P(A \cap B)}{P(B)}$$

Applying a little bit more algebra,

$$P\left(A \cap B\right) = P(A)P(B|A)$$

$$\text{and} \quad P\left(A \cap B\right) = P(B)P(A|B)$$

$$\text{and thus} \quad P(B)P(A|B) = P(A)P(B|A)$$

$$\text{and we get} \quad P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

Bayes rule *

Example (continued)

Probability of disease given symptoms

Suppose a TB test is 95% accurate
Suppose that TB is in 0.1% of population

What is P (TB | positive)?

Example (continued)

$$P(TB \mid positive)$$

$$= \frac{P(positive \mid TB) P(TB)}{P(positive)}$$

$$= \frac{P(positive \mid TB) P(TB)}{P(positive \mid TB) P(TB) + P(positive \mid \widetilde{TB}) P(\widetilde{TB})}$$

$$P(TB \mid positive)$$

$$= \frac{P(positive \mid TB) P(TB)}{P(positive)}$$

$$= \frac{P(positive \mid TB) P(TB)}{P(positive \mid TB) P(TB) + P(positive \mid \widetilde{TB}) P(\widetilde{TB})}$$

$$= \frac{(0.95)(0.001)}{(0.95)(0.001) + (0.05)(0.999)}$$

$$\cong 0.0187$$

Random Variables

Random variables
- Defined by functions mapping outcomes to values
- By choice, whatever we are interested in
- Typically denoted by uppercase letters (e.g., X)
- Generic values are corresponding lower case letters
- Shorthand: P(x) = P(X=x)
- Value “type” is arbitrary (typically categorical or real)

Example (from K&F)
- Outcomes are student grades (A,B,C)
- Random variable G=f_{GRADE}(student)
- $P(A) \equiv P(G = A) \equiv P(\{w \in \Omega : f_{GRADE}(w) = A\})$

Joint Distributions of Random Variables

Joint distribution of random variables

$$P(X,Y) \equiv P(X=x, Y=y) \equiv P(\{w \in \Omega : X(w)=x \text{ and } \mathrm{Y}(w)=y\})$$

Conditional definition, Bayes rule, chain rule all apply.

Marginal distributions ("sum rule")

$$P(X) = \sum_{Y} P(X,Y) \quad *$$

Chain (product) rule (two variable case of chain rule)

$$P(X,Y) = P(X \mid Y)P(Y) \quad *$$

Basic terminology and rules

Conditional probability

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

Bayes

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

$P(X \mid Y) \propto P(Y|X)P(X)$ (when Y is constant, i.e., evidence) *

Normalization

Often we will deal with quantities or functions which are proportional to probabilities (OK if we just want argmax)

To convert such quantities to probabilities we *normalize*.

$$\text{if} \quad p(x) \propto P(X=x) \quad \text{then} \quad P(X=x) = \frac{p(x)}{\sum_{x} p(x)}$$

Example: $P(X \mid Y) \propto P(X,Y)$

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

Independence

This can cause confusion. If P(Y) is zero, the other case cannot be used (divide by zero). However, in this case, Y never happens, and so we (a priori) have a choice to declare whether X is independent from Y or not. However, under scrutiny, the choice does make sense, and allows consistency with the second definition. Note that the second formula works in this (weird) case because if P(Y)=0, then P(X,Y) is also 0.

$$X \perp Y \quad \Leftrightarrow \quad P(X|Y) = P(X) \quad \text{or } \mathrm{P(Y)=0} \quad *$$

$$X \perp Y \quad \Leftrightarrow \quad P(X,Y) = P(X)P(Y) \quad *$$

Note that Bishop uses $\perp\!\!\!\perp$ instead of $\perp$

Conditional Independence

$X \perp Y \mid Z \quad \Leftrightarrow \quad P(X|Y,Z) = P(X \mid Z)$ or P(Y,Z)=0 *

Equivalent, sometimes more convenient definition

$X \perp Y \mid Z \quad \Leftrightarrow \quad P(X,Y|Z) = P(X \mid Z)P(Y \mid Z)$ *

Probabilistic Queries

Bold face because these are vectors of variables

Organize variables into

Evidence (observed), **E**

Query (what you want to know), **Y**

Hidden (leftover), **X** (for completeness)

Generic Query: P(**Y**|**E**)

This leads to a distribution over Y given the evidence

Note that X is marginalized out

We can use this to make a decision

Simplest is most probable, i.e., $\underset{\mathbf{Y}}{\text{Argmax}}\ P(\mathbf{Y},\mathbf{E})$

MAP Query (most probably configuration of variables):

$MAP(\mathbf{W} \mid \mathbf{E}) = \underset{\mathbf{W}}{\text{Argmax}}\ P(\mathbf{W},\mathbf{E}) \qquad (\mathbf{W}=\mathbf{Y} \cup \mathbf{X})$

Example

$P(x_1, y_2) = P(X= x_1 \text{ AND } Y= y_2)$

		Y	
		y_1	y_2
X	x_1	0.04	0.36
	x_2	0.30	0.30

$P(x_1)=P(x_1,y_1)+P(x_1,y_2)$
[i.e., sum across]

		Y		
		y_1	y_2	
X	x_1	0.04	0.36	0.4 ← $P(x_1)$
	x_2	0.30	0.30	0.6 ← $P(x_2)$
		0.34	0.66	P(x)

(Recall that P(x) is short hand for the probability that the random variable X takes the value x, similarly for P(y)).

		Y y_1		
X	x_1	0.04	0.04 / 0.34	$P(x \vert y_1)$
	x_2	0.30	0.30 / 0.34	
		P(0.34)		

(Recall that $P(x|y_1)$ is short hand for the probability that the random variable X takes the value x, given that the random variable Y has value y_1)

		Y y_1	y_2	
X	x_1	0.04	0.36	0.4
	x_2	0.30	0.30	0.6
		0.34	0.66	

Arg max P(x,y) is (x_1, y_2)

Arg max P(x) is (x_2)

Arg max P(y) is (y_2)

Arg max P(x,y) is **not necessarily** (Arg max P(x), Arg max P(y))

Discrete Distributions (Bernoulli)

$x \in \{0,1\}$ (e.g., 1 is "heads" and 0 is "tails")

$\mathrm{p}(x=1|\mu)=\mu$

$Bern(x \mid \mu)=\mu^{x}(1-\mu)^{(1-x)}$

Code for sampling a Bernoulli

```
a=rand()

if (a<u) return heads
else return tails
```

Discrete Distributions (Binomial)

How likely it is that we get m "heads" in N tosses?

$$Bin(m|N,\mu) = \binom{N}{m}\mu^m(1-\mu)^{N-m}$$

where $\binom{N}{m} \equiv \dfrac{N!}{(N-m)!m!}$

Discrete Distributions (Binomial)

Probability distribution for getting m "heads" in N tosses.

$$Bin(m|N,\mu) \;=\; \underbrace{\binom{N}{m}}_{\substack{\text{Number of}\\ \text{ways to get}\\ m \text{ heads}\\ \text{in } N \text{ tosses.}}} \bullet \underbrace{\mu^m(1-\mu)^{N-m}}_{\substack{\text{Probility of each}\\ \text{way to get } m \text{ heads}\\ \text{in } N \text{ tosses}}}$$

where $\binom{N}{m} \equiv \dfrac{N!}{(N-m)!m!}$

Example
N=3, m=2
HHT
HTH
THH

Multi-outcome Bernoulli

Simple extensions to Bernoulli to multiple outcomes (e.g., a six sided die).

Let K be the number of outcomes.

Now we use vectors for u and x, *i.e.*, $\mathbf{u}$ and $\mathbf{x}$.

$\mathbf{x}$ is a vector of 0's and exactly one 1 for observed outcome (e.g., rolling 3 with a 6 sided die is (0,0,1,0,0,0).

$$p(\mathbf{x} \mid \mathbf{u}) = \prod_{k=1}^{K} u_k^{x_k} \qquad \text{(note that } \sum_{k=1}^{K} u_k = 1\text{)}$$

Multinomial

Extension of binomial to multiple outcomes.

Let K be the number of outcomes.

$$Mult(m_1, m_2, \ldots, m_K) = \binom{N}{m_1 \;\; m_2 \;\; \ldots \;\; m_K} \prod_{k-1}^{K} \mu_k^{m_k}$$

where $\binom{N}{m_1 \;\; m_2 \;\; \ldots \;\; m_K} = \left(\dfrac{N!}{m_1! \;\; m_2! \;\; \ldots \;\; m_K!}\right)$

and $\sum_{k=1}^{K} m_k = N$

Continuous Spaces

Outcome space is observation of real values (e.g., height, mass)

Example, a random variable, X, can take any value in [0,1] with equal probability.

We say that X is uniformly distributed over [0,1].

Here, P(X=x) = 0 (uncountable number of possibilities).

To deal with this, we use Probability Density Functions.

Probability Density Functions

$p : \mathbb{R} \mapsto \mathbb{R}$ is a probability density function for X if $p(x) \geq 0$ and

$$\int_{Val(X)} p(x)dx = 1$$

$$P(a \leq X \leq b) = \int_a^b p(x)dx \qquad \text{(Probality of the event that x} \in \text{[a,b])}$$

$$P(X \in \Delta X) \cong p(x)|\Delta X| \qquad \text{(For small } \Delta X\text{)}$$

Note that P $\in [0,1]$ but $p(x)$ **can be larger** than 1.

Example one

A random variable is uniformly distributed between 0.4 and 0.6, and never occurs outside of that range.

$$p(x) = \begin{cases} \kappa & x \in [0.4, 0.6] \\ 0 & \text{otherwise} \end{cases}$$

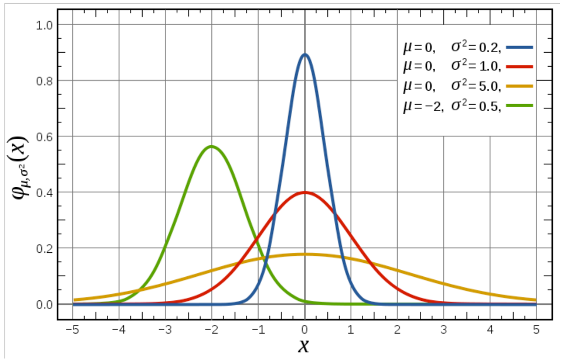
$$\int_{0.4}^{0.6} p(x)dx = \int_{0.4}^{0.6} \kappa \, dx = (0.2)\kappa = 1$$

$$\kappa = \frac{1}{0.2} = 5 \qquad \text{and thus} \qquad p(x) = \begin{cases} 5 & x \in [0.4, 0.6] \\ 0 & \text{otherwise} \end{cases}$$

Example two

The univariate Gaussian (or Normal) distribution

$$\mathbb{N}(\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Joint Density Functions

Analogous to univariate case (illustrated with two variables)

$$\iint_{Val(X)\times Val(Y)} p(x,y)\,dx\,dy = 1$$

$$P(a_X \le X \le b_X,\ a_Y \le Y \le b_Y) = \int_{a_Y}^{b_Y}\int_{a_X}^{b_X} p(x,y)\,dx\,dy$$

Example--- multivariate Gaussian

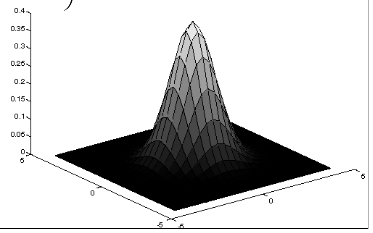
$$\mathbb{N}(\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{k}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}}\exp\left(\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\,\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

k is the number of variables (dimension)

If the variables are independent, then the covariance is diagonal

$$\mathbb{N}(\boldsymbol{\mu},\boldsymbol{\sigma^2}) = \frac{1}{(2\pi)^{\frac{k}{2}}\prod_{i=1}^{k}\sigma_i}\exp\left(\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\left(diag(\sigma^2)\right)^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

$$= \prod_{i=1}^{k}\mathbb{N}(\mu_i,\sigma_i^2)$$



Marginalization

$$p(x) = \int_{-\infty}^{\infty} p(x,y)\,dy$$

Conditional Distributions

$$p(y\,|\,x) = \frac{p(x,y)}{p(x)} \quad \text{where } p(x)\neq 0$$

Can get this by marginalizing

$$p(x) = \int_{-\infty}^{\infty} p(x,y)\,dy$$

Gaussian Facts

For a multivariate Gaussian p($\mathbf{x}_a$, $\mathbf{x}_b$) with variables partitioned into $\mathbf{x}_a$ and $\mathbf{x}_b$ we have:

p($\mathbf{x}_a$) is also Gaussian

and

p($\mathbf{x}_a \mid \mathbf{x}_b$) is also Gaussian

Chapter 2.3 of Bishop has a very thorough treatment of the Gaussian distribution.

Expectation

$$E_p[X] = \sum_x x \cdot P(x) \qquad \text{(discrete)}$$

$$E_p[X] = \int x \cdot p(x)\, dx \qquad \text{(continuous)}$$

$$E_p[X+Y] = E_p[X] + E_p[Y]$$

Implicit definition of a new random variable

Variance

Recall that this is our symbol for independent.

$$Var(X) = E_p\left[\left(X - E_p[X]\right)^2\right]$$

$$Var(X+Y) = Var(X) + Var(Y) \qquad \text{(when X $\perp$ Y)}$$

$$Var(aX) = a^2 \cdot Var(X)$$

Standard deviation, $\sigma_X = \sqrt{Var(X)}$

Sampling Continuous Distributions

- Suppose you want to generate samples from (i.e., simulate a probability distribution).
- The typical tool at your disposal is a pseudo random number generator returning approximately uniformly distributed rational numbers in [0,1]
- Sampling Bernoulli processes is straightforward
- Variants of uniform distributions are also easy
- Example: $p(x) = \begin{cases} 5 & x \in [0.4, 0.6] \\ 0 & \text{otherwise} \end{cases}$

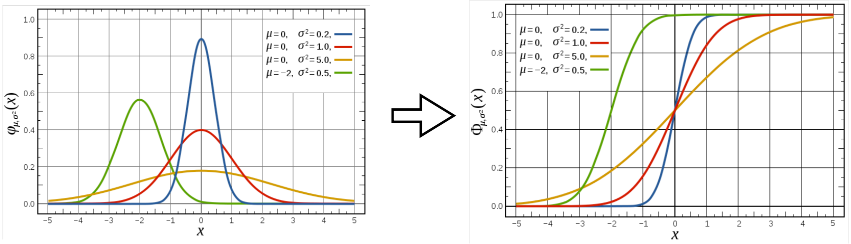
Sampling Continuous Distributions

- N(0,1) is less obvious (there are standard fast methods)
- A general approach for sampling a continuous distribution (sometimes call inverse transformation sampling) is based on the cumulative distribution function, CDF, denoted by F(x)

Cumulative Distribution Function

$$F(x) = P(X \leq x)$$

$$= \int_{-\infty}^{x} p(x)dx \quad \text{(continuous distributions)}$$



Sampling Continuous Distributions

We know how to sample y uniformly from [0,1]

We want to map $y \Rightarrow x \in [-\infty,\infty]$ where is x distributed as $p(x)$

For simplicity, map them monotonically (bigger $y \Rightarrow$ bigger x)

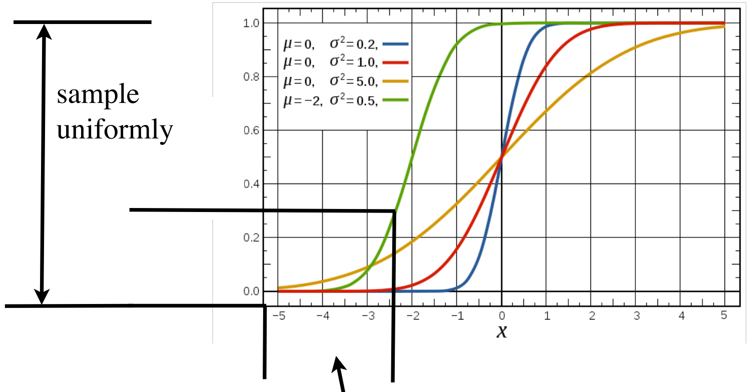
All samples in U=[0,y] should map to total probability y over $p(x)$.

We know how to sample y uniformly from [0,1]

We want to map $y \Rightarrow x \in [-\infty,\infty]$ where is x distributed as $p(x)$

For simplicity, map them monotonically (bigger $y \Rightarrow$ bigger x)

All samples within U=[0,y] should map to total probability y from $p(x)$.



sample uniformly

Accounts for 0.3 probability mass

Sampling Continuous Distributions

We know how to sample y uniformly from [0,1]

We want to map $y \Rightarrow x \in [-\infty,\infty]$ where is x distributed as $p(x)$

For simplicity, map them monotonically (bigger $y \Rightarrow$ bigger x)

All samples in U=[0,y] should map to total probability y in $p(x)$

So U=[0,y] maps into $P=[-\infty,x]$, where y=$\int_{-\infty}^{x} p(x')\,dx' = F(x)$

Further, a sample y $\in [0,1]$ should map to x such that $y = F(x)$

In other words, $x = F^{-1}(y)$

Sampling Continuous Distributions

- To sample a distribution p(x) (crude instructional algorithm)

```
Prepare and approximation of $F(x)$
in a vector F=($x_1$, $x_2$, $x_3$, ... , $x_N$)

Loop
  sample $y \in [0,1]$
  find i so that F($x_i$)<y and F($x_{i+1}$)>y
  report ($x_i + x_{i+1}$)/2
```

Example (from Bishop, PRML)

Estimating the mean of a univariate Gaussian

Assume that the variance is known.

Given data points x_i, what is the "best" estimate for the mean?

$$p(u|\{x_i\}) \propto p(\{x_i\}|u) \quad \text{(assuming uniform prior)}$$

$$p(\{x_i\}|u) = \prod_i p(x_i|u)$$

$$\propto \prod_i e^{-\frac{(x_i-u)^2}{2\sigma^2}}$$

We can maximize the probility by minimizing the negative log

$$-\log\left(\prod_i e^{-\frac{(x_i-u)^2}{2\sigma^2}}\right) \propto \sum_i (x_i-u)^2$$

$$u_{ML} = \arg\max_u\left(\sum_i (x_i-u)^2\right)$$

Differentiating and setting to zero reveals that

$$u = \frac{1}{N}\sum x_i$$

Example (from Bishop, PRML)

Estimating the mean of a univariate Gaussian

Assume that the variance is known.

Given data points x$_i$, what is the "best" estimate for the mean?

The maximum likelihood estimate is $\mu_{ML} = \frac{1}{N}\sum_i x_i$

But what if the number of points is small?

Lets consider the case where we want to incorporate prior information.

IE, let's do Bayes.

$$
\begin{aligned}
p\left(\mu \mid \{x_i\}\right) &\propto p(\mu)p(\{x_i\} \mid \mu) \\
&= p(\mu)\prod_i p(\{x_i\} \mid \mu) \\
&\propto p(\mu)\prod_i \exp\left(-\frac{\left(x_i-\mu\right)^2}{2\sigma^2}\right)
\end{aligned}
$$

What should we use for $p(\mu)$?

$$p\left(\mu \mid \{x_i\}\right) \propto p(\mu)\prod_i \exp\left(-\left(x_i-\mu\right)^2\right)$$

By inspection, if $p(\mu) \propto \exp\left(-\left(\mu_0-\mu\right)^2\right)$ then

the form of the posterior is the same as the prior.

IE, given known variance, a conjugate prior for the mean of the Gaussian is a Gaussian.

Conjugacy is convenient for several reasons, but one motivating observation is Bayesian updating whereby yesterday's posterior is used for today's prior.

Quick aside one (Bayesian update)

Consider two successive groups of observations that are conditionally independent given the model

$$
\begin{aligned}
p(\theta, \mathbf{x}_2, \mathbf{x}_1) &= p\left(\mathbf{x}_2 \middle| \theta\right) p\left(\mathbf{x}_1 \middle| \theta\right) p\left(\theta\right) \\
&= p\left(\mathbf{x}_2 \middle| \theta\right) p\left(\theta \middle| \mathbf{x}_1\right) p\left(\mathbf{x}_1\right)
\end{aligned}
$$

SO

$$p(\theta, \mathbf{x}_2 | \mathbf{x}_1) = p\left(\mathbf{x}_2 \middle| \theta\right) \underbrace{p\left(\theta \middle| \mathbf{x}_1\right)}_{\substack{\text{updated prior,} \\ \text{after seeing } \mathbf{x}_1}}$$

Quick aside two (Conjugacy)

Informal definition: Given a likelihood function

$l(\theta$,x)=p(x|θ)   (we reverse θ and x when we call it a likelihood function)

a (prior) distribution is natural distribution where the posterior,

$p(\theta \mid x) \propto p(x \mid \theta)p(\theta)$, has the same form as p(θ).

Back to our problem.

$$p\left(\mu \mid \{x_i\}\right) \propto \underbrace{\exp\left(-\frac{\left(\mu_0-\mu\right)^2}{\sigma_0^2}\right)}_{\text{conjugate prior for the likelihood}} \underbrace{\prod_i \exp\left(-\frac{\left(x_i-\mu\right)^2}{\sigma^2}\right)}_{\text{likelihood}}$$

To find the MAP (maximum a posteriori) estimate, we maximize.

Maximizing is the same as minimizing the negative log.

$$-\log\left(p\left(\mu \mid \{x_i\}\right)\right)=\frac{\left(\mu_0-\mu\right)^2}{\sigma_0^2}+\sum_i \frac{\left(x_i-\mu\right)^2}{\sigma^2}$$

differentiating and setting derivatives to zero gives

$$\frac{\mu_0}{\sigma_0^2}+\frac{1}{\sigma^2}\sum_i x_i=\frac{\mu}{\sigma_0^2}+\frac{N\mu}{\sigma^2}$$

$$-\log\left(p\left(\mu \mid \{x_i\}\right)\right)=\frac{\left(\mu_0-\mu\right)^2}{\sigma_0^2}+\sum_i \frac{\left(x_i-\mu\right)^2}{\sigma^2}$$

differentiating and setting derivatives to zero gives

$$\frac{\mu_0}{\sigma_0^2}+\frac{1}{\sigma^2}\sum_i x_i=\frac{\mu}{\sigma_0^2}+\frac{N\mu}{\sigma^2}$$

algebra reveals that

$$\mu_{MAP}=\frac{\frac{\mu_0}{\sigma_0^2}+\frac{N}{\sigma^2}\mu_{ML}}{\frac{1}{\sigma_0^2}+\frac{N}{\sigma^2}}=\frac{\frac{\mu_0}{\sigma_0^2}}{\frac{1}{\sigma_0^2}+\frac{N}{\sigma^2}}+\frac{\frac{N}{\sigma^2}\mu_{ML}}{\frac{1}{\sigma_0^2}+\frac{N}{\sigma^2}}=\frac{\sigma^2}{\sigma^2+\sigma_0^2 N}\mu_0+\frac{N\sigma_0^2}{\sigma^2+N\sigma_0^2}\mu_{ML}$$

Example (from Bishop, PRML)

Unknown variance or mean and variance

Similar stories can be told if the mean is known and the variance is not, or both are unknown. We will only set up the problem to have a look at the conjugate priors.

Simplify things by using the inverse of the covariance matrix which is called the precision matrix.

In the univariate case this is simply: $\lambda = \frac{1}{\sigma^2}$

Example (from Bishop, PRML)

Known mean, unknown variance

$$p(\{x_i\} \mid \lambda) = \prod_{i=1}^{N} \mathbb{N}\left(x_i \mid \mu, 1/\lambda\right)$$

$$= \prod_{i=1}^{N} \left\{ \left(\frac{\lambda}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{\lambda}{2}(x_i - \mu)^2\right) \right\} \qquad (u \text{ is constant})$$

$$\propto \lambda^{N/2} \exp\left\{-\frac{\lambda}{2} \sum_i (x_i - \mu)^2 \right\}$$

constant

Inspection reveals that multiplying this by a gamma distribution

$$\mathrm{Gam}(\lambda \mid a, b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda)$$

yields a posterior of the same form. The normalization constant, $\Gamma(a)$ is the "gamma" function, which extends the concept of factorial to real numbers. $\Gamma(n) = (n-1)!$, for postive integers n. Also $\Gamma(x+1) = x\Gamma(x)$ for postive reals.

"Inspection"

$$p(\{x_i\} \mid \lambda) \propto \lambda^{N/2} \exp\left\{-\frac{\lambda}{2} \sum_i (x_i - \mu)^2 \right\} = \lambda^{N/2} \exp\left\{-\frac{\lambda}{2} K\right\}$$

$$\mathrm{Gam}(\lambda \mid a, b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda) \propto \lambda^{a-1} \exp(-b\lambda)$$

$$p(\{x_i\} \mid \lambda)\,\mathrm{Gam}(\lambda \mid a, b) \propto \lambda^{N/2} \lambda^{a-1} \exp\left\{-\frac{\lambda}{2} K\right\} \exp\{-b\lambda\}$$

$$= \lambda^{((N/2)+a-1)} \exp\left\{-\lambda\left(\left(K/2\right)+b\right)\right\}$$

Gamma distribution illustrated (*)

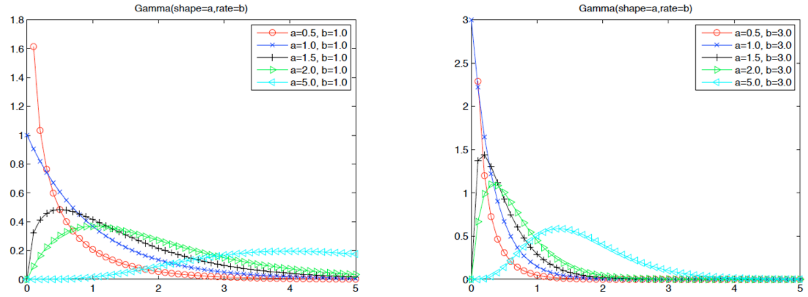


Figure 1: Some $Ga(a, b)$ distributions. If $a < 1$, the peak is at 0. As we increase b, we squeeze everything leftwards and upwards. Figures generated by `gammaDistPlot2`.

* From an on-line note by Kevin Murphy (*www.cs.ubc.ca/~murphyk/Teaching/CS340-Fall07/reading/NG.pdf*)

Example (from Bishop, PRML)

Unknown mean and variance

$$
\begin{aligned}
p(u,\lambda) &= p(u \mid \lambda)p(\lambda) \\
&= N\left(u \middle| u_o, (\beta\lambda)^{-1}\right) Gam(\lambda \mid a,b)
\end{aligned}
$$

Here a,b,β are constants. This is the normal-gamma (Gaussian-gamma) distribution.

(Derivation follows for completeness)

◆◆

Example (from Bishop, PRML)

Unknown mean and variance

Indicates optional material

$$
\begin{aligned}
p(\{x_i\} \mid \lambda) &= \prod_{i=1}^{N} \mathbb{N}\left(x_i \mid \mu, {}^{1}\!/\!_{\lambda}\right) \\
&= \prod_{i=1}^{N} \left\{ \left(\frac{\lambda}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{\lambda}{2}(x_i-\mu)^2\right) \right\} \\
&\propto \lambda^{N/2} \exp\left\{ -\frac{\lambda}{2}\sum_i (x_i-\mu)^2 \right\} \\
&= \lambda^{N/2} \exp\left\{ -\frac{\lambda}{2}\sum_i x_i^2 + \lambda\mu\sum_i x_i - \frac{N\lambda}{2}\mu^2 \right\}
\end{aligned}
$$

(u is variable)

◆◆

$$
\begin{aligned}
p(\{x_i\} \mid u,\lambda) &\propto \lambda^{N/2} \exp\left\{ -\frac{\lambda}{2}\sum_i x_i^2 + \lambda\mu\sum_i x_i - \frac{N\lambda}{2}\mu^2 \right\} \\
&= \lambda^{N/2} \left(\exp\left(-\frac{\lambda\mu^2}{2}\right) \right)^{N} \exp\left\{ \lambda\mu\sum_i x_i - \frac{\lambda}{2}\sum_i x_i^2 \right\} \\
&= \lambda^{N/2} \left(\exp\left(-\frac{\lambda\mu^2}{2}\right) \right)^{N} \exp\left(C\lambda\mu - D\lambda\right)
\end{aligned}
$$

◆◆

From the previous slide

$$
p(\{x_i\} \mid u,\lambda) \propto \lambda^{N/2} \left(\exp\left(-\frac{\lambda\mu^2}{2}\right) \right)^{N} \exp\left(\overset{\sum_i x_i}{\overset{\updownarrow}{C}}\lambda\mu - \overset{\frac{1}{2}\sum_i x_i^2}{\overset{\updownarrow}{D}}\lambda\right)
$$

So a conjugate prior of the form

$$
p(u,\lambda) \propto \lambda^{\beta/2} \left(\exp\left(-\frac{\lambda\mu^2}{2}\right) \right)^{\beta} \exp\left(c\lambda\mu - d\lambda\right)
$$

will do (recall that exp(x)•exp(y) = exp(x+y)).

We now manipulate the formula to a more standard form.

$$p(u,\lambda) \propto \lambda^{\beta/2}\left(\exp(-\frac{\lambda\mu^2}{2})\right)^{\beta}\exp(c\lambda\mu - d\lambda)$$

$$= \lambda^{\beta/2}\left(\exp(-\frac{\lambda\beta}{2}\mu^2)\right)\exp(c\lambda\mu - d\lambda)$$

$$= \lambda^{\beta/2}\left(\exp(-\frac{\lambda\beta}{2}\mu^2 + c\lambda\mu - d\lambda)\right)$$

$$= \lambda^{\beta/2}\left(\exp\left(-\frac{\lambda\beta}{2}\left(\mu^2 - \frac{2c\mu}{\beta} + \frac{2d}{\beta}\right)\right)\right)$$

From the previous slide

$$p(u,\lambda) \propto \lambda^{\beta/2}\left(\exp\left(-\frac{\lambda\beta}{2}\left(\mu^2 - \frac{2c\mu}{\beta} + \frac{2d}{\beta}\right)\right)\right)$$

$$\mu^2 - \left(\frac{2c}{\beta}\right)\mu + \frac{2d}{\beta} = \left(\mu - \frac{c}{\beta}\right)^2 + \frac{2d}{\beta} - \frac{c^2}{\beta^2}$$

$$p(u,\lambda) \propto \lambda^{\beta/2}\exp\left(-\frac{\lambda\beta}{2}\left(\mu - \frac{c}{\beta}\right)^2\right)\exp\left(-\lambda\left(d - \frac{c^2}{2\beta}\right)\right)$$

$$= \exp\left(-\frac{\lambda\beta}{2}\left(\mu - \frac{c}{\beta}\right)^2\right)\lambda^{\beta/2}\left(\exp\left(-\lambda\left(d - \frac{c^2}{2\beta}\right)\right)\right)$$

From the previous slide

$$p(u,\lambda) \propto \exp\left(-\frac{\lambda\beta}{2}\left(\mu - \frac{c}{\beta}\right)^2\right)\lambda^{\beta/2}\left(\exp\left(-\lambda\left(d - \frac{c^2}{2\beta}\right)\right)\right)$$

$$\propto \mathbb{N}\left(\mu \mid \mu_0, (\lambda\beta)^{-1}\right) Gam(\lambda \mid a,b)$$

where $\mu_0 = \dfrac{c}{\beta}$ and $a = 1 + \dfrac{\beta}{2}$ (*) and $b = d - \dfrac{c^2}{2\beta}$

Recall that $Gam(\lambda \mid a,b) \propto \lambda^{a-1}\exp(-b\lambda)$

$$p(\mu,\lambda) = p(\mu \mid \lambda)\,p(\lambda) = \mathbb{N}\left(\mu \mid \mu_0, (\lambda\beta)^{-1}\right) Gam(\lambda \mid a,b)$$

This is called the Gaussian-Gamma function

*According to Bishop, this is how the gamma parameter, *a*, relates to the Gaussian variance scale beta, but the powers of lambda from the normal do not seem to be accounted for --- regardless, the conjugate formula is still correct.

Beta (and Dirichlet) distributions

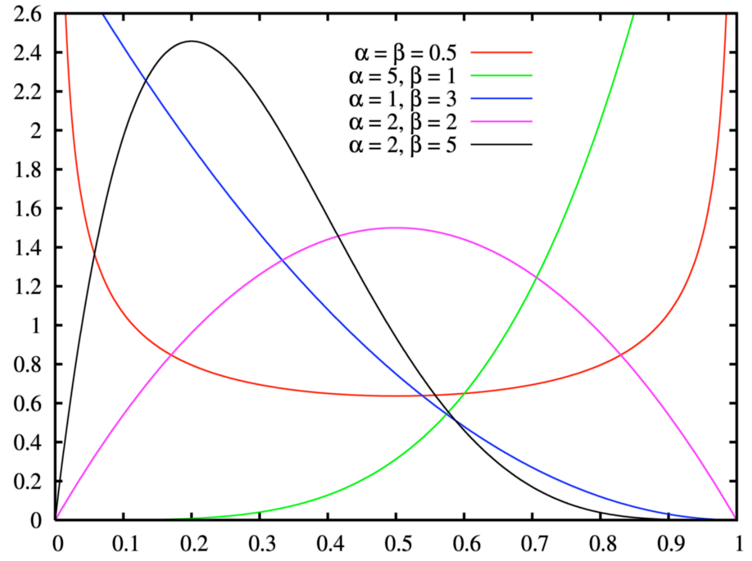
Beta (binary case)

Conjugate prior for the Bernoulli and binomial distributions

Dirichlet (multi-outcome case)

Conjugate priors for the multi outcome Bernoulli and multinomial distributions

$$Beta(u \mid a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} u^{a-1} (1-u)^{b-1}$$



$$Beta(u \mid a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} u^{a-1} (1-u)^{b-1}$$

$$Bern(x \mid \mu) = \mu^{x} (1-\mu)^{(1-x)}$$

(You should be able to tell the rest of the story ...)

More on priors

If we leave off the prior, then we are completely ignorant.

Note that the prior might be the uniform distribution over all numbers

This is not a PDF!

Such priors are called improper.

A more interesting example is p(k)=1/k.

Everything can work out fine if the posterior is a PDF.

Bayesian Sequential Update

- For independent sequential events

$$p\left(\theta \mid D_{1:N}\right)=\left\{p(\theta)\prod_{i=1}^{N-1}\left(p\left(D_i \mid \theta\right)\right)\right\}p\left(D_N \mid \theta\right)$$

New prior

Already introduced with the example for the Bayesian estimate of the mean

Predictive Distribution

- The general predictive distribution marginalizes over uncertain model parameters

$$p\left(x \mid X\right)=\int p(x \mid \theta)p(\theta \mid X)d\theta$$

Test data

Training data

Bayesian statistics summary

- Bayesian statistical models
  - We prefer generative models for likelihood (and prior)
  - Conjugate priors are useful
  - Bayesian updating for independent sequence of data
- Inference uses Bayes rule to “invert” the forward model
- Predictive distribution
  - Marginalizes out uncertainty about models
- Related topics coming soon
  - Model selection
  - Decision making