Announcements

Assignment two due Friday, Feb 10.

Office hour planned for Friday.

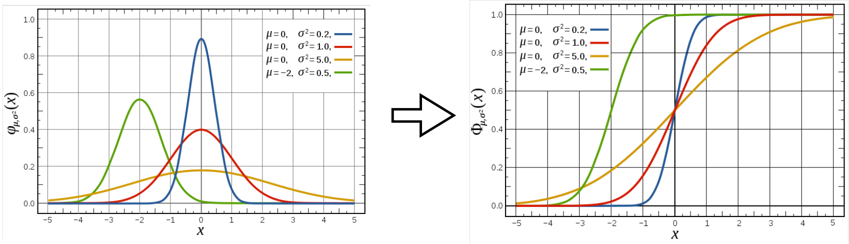
Next week:
1) I am away Wednesday through Friday
2) Thursday class will be taken by Kyle
3) No Friday office hour

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

$$
\begin{aligned}
F(x) &= P(X \leq x) \\
&= \int_{-\infty}^{x} p(x)\,dx \quad \text{(continuous distributions)}
\end{aligned}
$$



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)

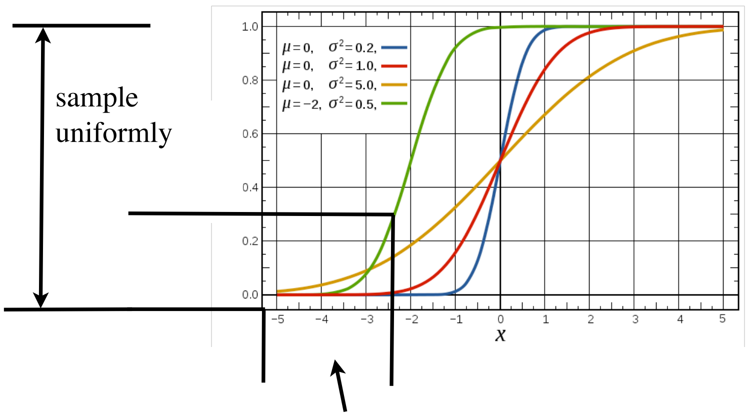
All samples in U=[0,y] should map to total probability y over $p(x)$.

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 $F=(x_1, x_2, x_3, \dots, x_N)$

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?

$$p\left(u \middle| \{x_i\}\right) \propto p\left(\{x_i\} \middle| u\right) \qquad \text{(assuming uniform prior)}$$

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

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

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.

$$p\left(\mu \mid \{x_i\}\right) \propto p(\mu)p(\{x_i\} \mid \mu)$$

$$= p(\mu)\prod_i p(\{x_i\} \mid \mu)$$

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

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

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

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

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

so

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

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.

$$p\left(\mu \mid \{x_i\}\right) \propto \exp\left(-\frac{\left(\mu_0-\mu\right)^2}{\sigma_0^2}\right)\prod_i \exp\left(-\frac{\left(x_i-\mu\right)^2}{\sigma^2}\right)$$

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

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

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: $\lambda = \dfrac{1}{\sigma^2}$

Example (from Bishop, PRML)

Estimating the variance

$$p\left(\left\{x_i\right\} \mid \lambda\right) = \prod_i N\left(x_i \mid \mu, \lambda\right)$$

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

Inspection reveals that multiplying this by a gamma distribution

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

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.