CSC 2541: Bayesian Methods for Machine Learning

Radford M. Neal, University of Toronto, 2011

Lecture 1

Tentative List of Topics by Week

1:	Introduction to Bayesian inference
	Analytical inference with conjugate priors
2 & 3:	Monte Carlo methods for Bayesian inference
	Linear regression, logistic regression
4 & 5:	Mixture models, finite and infinite
	Latent variables (statistical and computational aspects)
	Variational methods for Bayesian inference
	Models using the Indian Buffet Process
6 & 7:	Gaussian process models for regression and classification
	Matrix computations
	Monte Carlo and other methods of inference
8 & 9:	Other models: neural networks, ??
10:	Test
11 & 12:	Project presentations

Evaluation

Five small exercises (1 week):	25%
Major assignment (4 weeks):	25%
Test:	20%
Project (individual or group):	30%

Projects may be done individually, in a group of two, or in a group of three for good reason (with special permission).

Tentative Dates

Test:	Mar 21
Major assignment:	Handed out Jan 31, due Feb 28
Project:	Proposal Feb 21 (emailed), Presentations Mar 28 & Apr 4 Report Apr 7

Introduction to Bayesian Inference

The Bayesian Approach to Machine Learning (Or Anything)

1) We formulate our knowledge about the situation probabilistically:
   - We define a *model* that expresses qualitative aspects of our knowledge (eg, forms of distributions, independence assumptions). The model will have some unknown *parameters*.
   - We specify a *prior* probability distribution for these unknown parameters that expresses our beliefs about which values are more or less likely, before seeing the data.
2) We gather data.
3) We compute the *posterior* probability distribution for the parameters, given the observed data.
4) We use this posterior distribution to:
   - Reach scientific conclusions, properly accounting for uncertainty.
   - Make predictions by averaging over the posterior distribution.
   - Make decisions so as to minimize posterior expected loss.

Finding the Posterior Distribution

The *posterior distribution* for the model parameters given the observed data is found by combining the prior distribution with the likelihood for the parameters given the data.

This is done using *Bayes' Rule*:

$$P(\text{parameters} \,|\, \text{data}) \quad = \quad \frac{P(\text{parameters})\, P(\text{data} \,|\, \text{parameters})}{P(\text{data})}$$

The denominator is just the required normalizing constant, and can often be filled in at the end, if necessary. So as a proportionality, we can write

$$P(\text{parameters} \,|\, \text{data}) \quad \propto \quad P(\text{parameters})\, P(\text{data} \,|\, \text{parameters})$$

which can be written schematically as

$$\text{Posterior} \quad \propto \quad \text{Prior} \;\times\; \text{Likelihood}$$

We make predictions by integrating with respect to the posterior:

$$P(\text{new data} \,|\, \text{data}) \quad = \int_{\text{parameters}} P(\text{new data} \,|\, \text{parameters})\, P(\text{parameters} \,|\, \text{data})$$

Representing the Prior and Posterior Distributions by Samples

The complex distributions we will often use as priors, or obtain as posteriors, may not be easily represented or understood using formulas.

A very general technique is to represent a distribution by a *sample* of many values drawn randomly from it. We can then:

- *Visualize* the distribution by viewing these sample values, or low-dimensional projections of them.
- Make *Monte Carlo* estimates for probabilities or expectations with respect to the distribution, by taking averages over these sample values.

Obtaining a sample from the prior is often easy. Obtaining a sample from the posterior is usually more difficult — but this is nevertheless the dominant approach to Bayesian computation.

Inference at a Higher Level: Comparing Models

So far, we've assumed we were able to start by making a definite choice of model. What if we're unsure which model is right?

We can compare models based on the *marginal likelihood* (aka, the *evidence*) for each model, which is the probability the model assigns to the observed data. This is the normalizing constant in Bayes' Rule that we previously ignored:

$$P(\text{data} \mid M_1) \;=\; \int_{\text{parameters}} P(\text{data} \,|\, \text{parameters}, M_1)\, P(\text{parameters} \,|\, M_1)$$

Here, M_1 represents the condition that model M_1 is the correct one (which previously we silently assumed). Similarly, we can compute $P(\text{data} \mid M_2)$, for some other model (which may have a different parameter space).

We might choose the model that gives higher probability to the data, or average predictions from both models with weights based on their marginal likelihood, multiplied by any prior preference we have for M_1 versus M_2.

A Simple Example — A Hard Linear Classifier

The problem:

We will be observing pairs $(x^{(i)}, y^{(i)})$, for $i = 1, \ldots, n$, where $x = (x_1, x_2)$ is a 2D "input" and y is a $-1/+1$ class indicator. We are interested in predicting y from x. We are not interested in predicting x, and this may not even make sense (eg, we may determine the $x^{(i)}$ ourselves).

Our informal beliefs:

We believe that there is a line somewhere in the input space that determines y perfectly — with -1 on one side, $+1$ on the other.

We think that this line could equally well have any orientation, and that it could equally well be positioned anywhere, as long as it is no more than a distance of three from the origin at its closest point.

We need to translate these informal beliefs into a *model* and a *prior*.

Formalizing the Model

Our model can be formalized by saying that

$$P(y^{(i)} = y \mid x^{(i)}, u, w) \quad = \quad \begin{cases} 1 & \text{if } y\,u\,(w^T x^{(i)} - 1) > 0 \\ 0 & \text{if } y\,u\,(w^T x^{(i)} - 1) < 0 \end{cases}$$

where $u \in \{-1, +1\}$ and $w = (w_1, w_2)$ are unknown *parameters* of the model. The value of w determines a line separating the classes, and u says which class is on which side. (Here, $w^T x$ is the scalar product of w and x.)

This model is rather dogmatic — eg, it says that y is **certain** to be $+1$ if $u = +1$ and $w^T x$ is greater than 1. A more realistic model would replace the probabilities of 0 and 1 above with ϵ and $1 - \epsilon$ to account for possible unusual items, or for misclassified items. ϵ might be another unknown parameter.

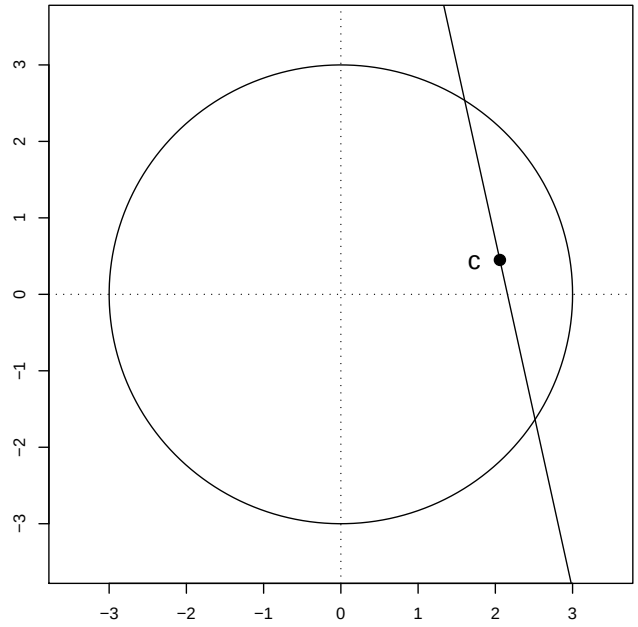
Formalizing the Prior

A line is completely determined by giving the point, c, on the line that is closest to the origin.

To formalize our prior belief that the line separating classes could equally well be anywhere, as long as it is no more than a distance of three from the origin, we decide to use a uniform distribution for c over the circle with radius 3.

Given c, we can compute $w = c/||c||^2$, which makes $w^T x = 1$ for points on the line. (Here, $||c||^2$ is the squared norm, $c_1^2 + c_2^2$.)

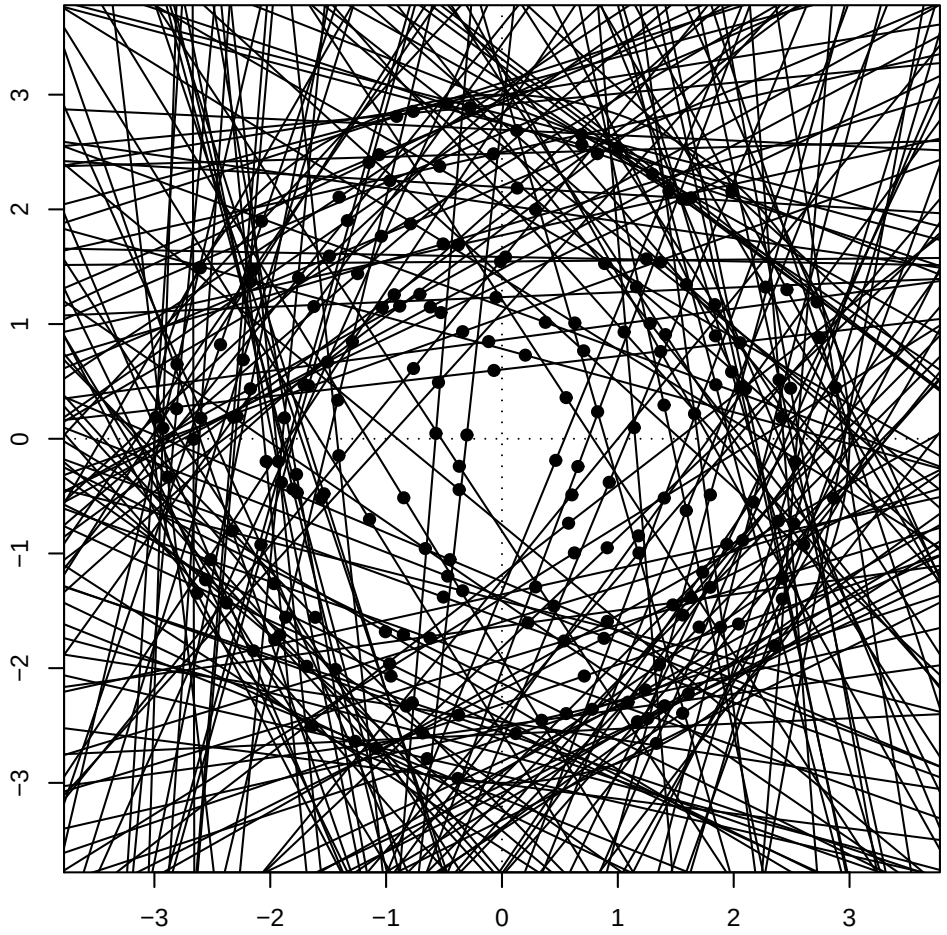
Here's an example:



We also say that u is equally likely to be $+1$ or -1, independently of w.

Looking at the Prior Distribution

We can check this prior distribution by looking at many lines sampled from it:

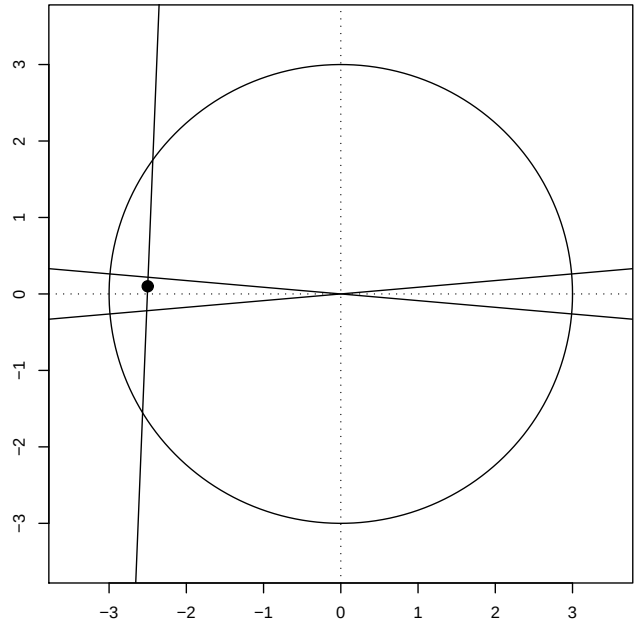


Something's wrong here. We meant for the lines to be uniformly distributed, but we see a sparse region near the origin.

Why This Prior Distribution is Wrong

Our first attempt at formalizing our prior beliefs didn't work. We can see why if we think about it.

Imagine moving a line that's within five degrees of vertical from left to right:

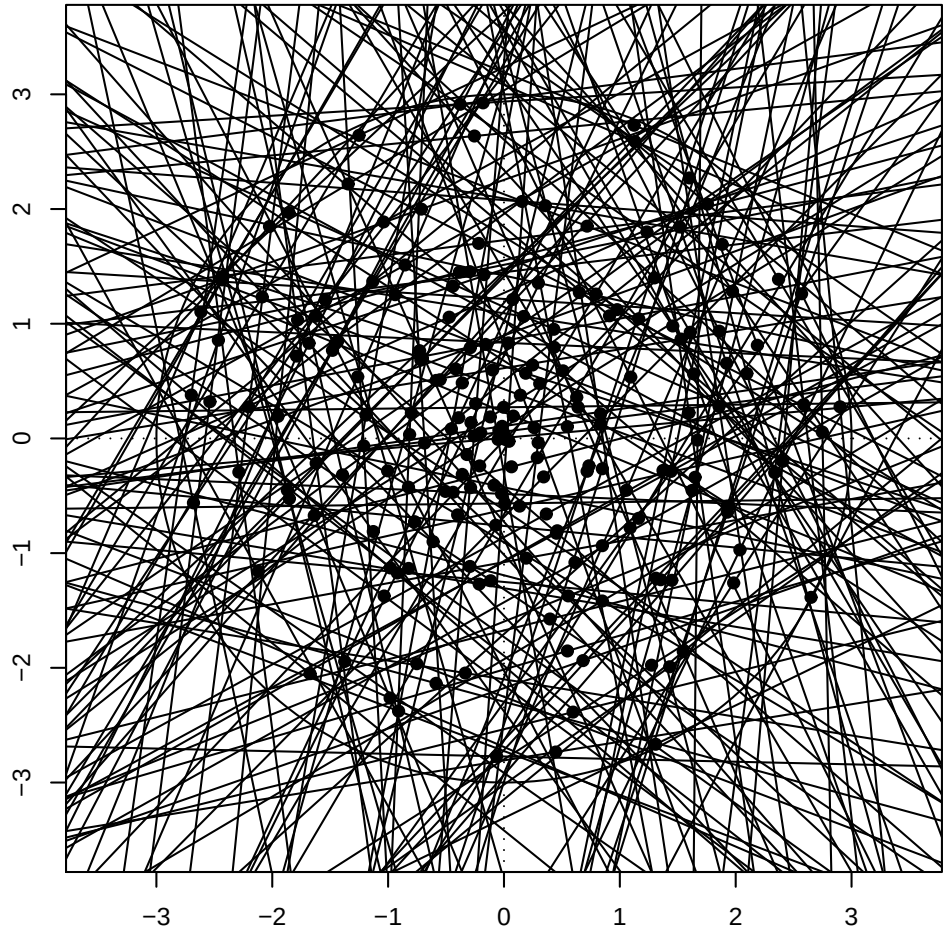


To stay within five degrees of vertical, the closest point to the origin has to be within the wedge shown. This becomes less and less likely as the origin is approached. We don't get the same probability of a near-vertical line for all horizontal positions.

Fixing the Prior Distribution

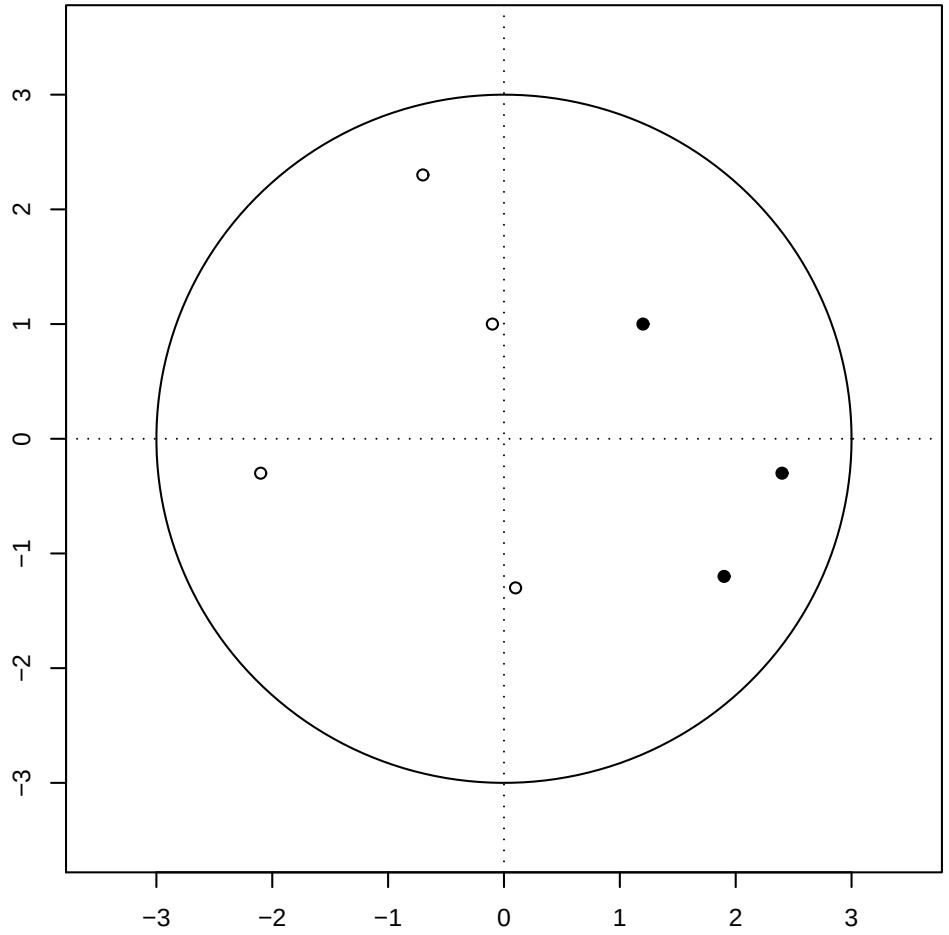
We can fix the prior by letting the closest point on the line to the origin be $c = rv$, with r uniformly distributed over $(0, 3)$ and v uniformly distributed over the unit circle.

Now a sample drawn from the prior looks the way we want it to:



Some Data Points

Now that we have defined our model and prior, let's get some data:



The black points are in class $+1$, the white points in class -1.

Posterior Distribution for the Hard Linear Classifier

For the hard linear classifier, the likelihood is either 0 or 1:

$$P(y^{(1)}, \ldots, y^{(n)} \mid x^{(1)}, \ldots, x^{(n)}, u, w) = \prod_{i=1}^{n} P(y^{(i)} \mid x^{(i)}, u, w)$$

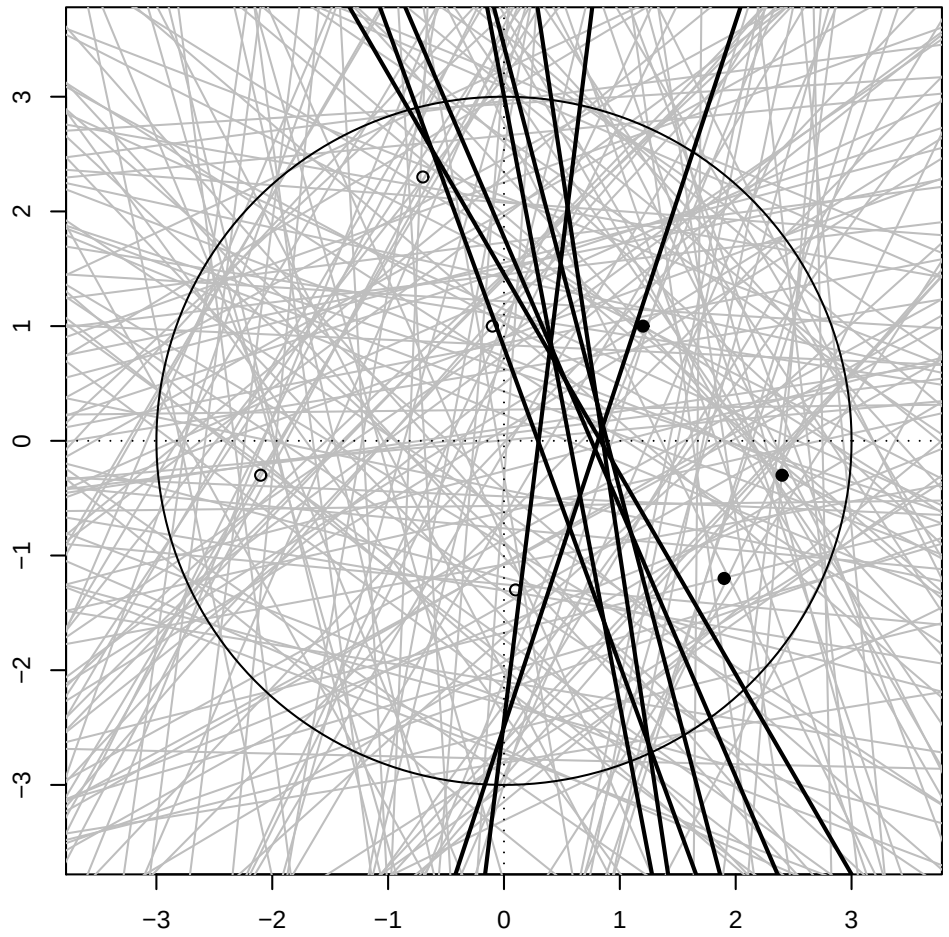
$$= \begin{cases} 1 & \text{if } y^{(i)}\, u\, (w^T x^{(i)} - 1) > 0, \text{ for } i = 1, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

The posterior distribution for u and w is therefore the same as their prior distribution, except that parameter values incompatible with the data are eliminated.

After renormalizing so that posterior probabilities integrate to one, the parameter values compatible with the data will have higher probability than they did in the prior.

Obtaining a Sample from the Posterior Distribution

To obtain a sample of values from the posterior, we can sample w values from the prior, but retain only those that are compatible with the data (for some u). Here's what we get using a sample of size 200:



The eight bold lines are a random sample from the posterior distribution.

Making a Prediction for a Test Case

The Bayesian predictive probability that in a test case with inputs x^*, the class, y^*, will be $+1$ is found by integrating/summing over the parameters w and u:

$$\begin{aligned} & P(y^* = +1 \,|\, x^*, (x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})) \\ & \quad = \int \sum_{u=\pm 1} P(y^* = +1 \,|\, x^*, u, w)\, P(u, w \,|\, x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})\, dw \end{aligned}$$

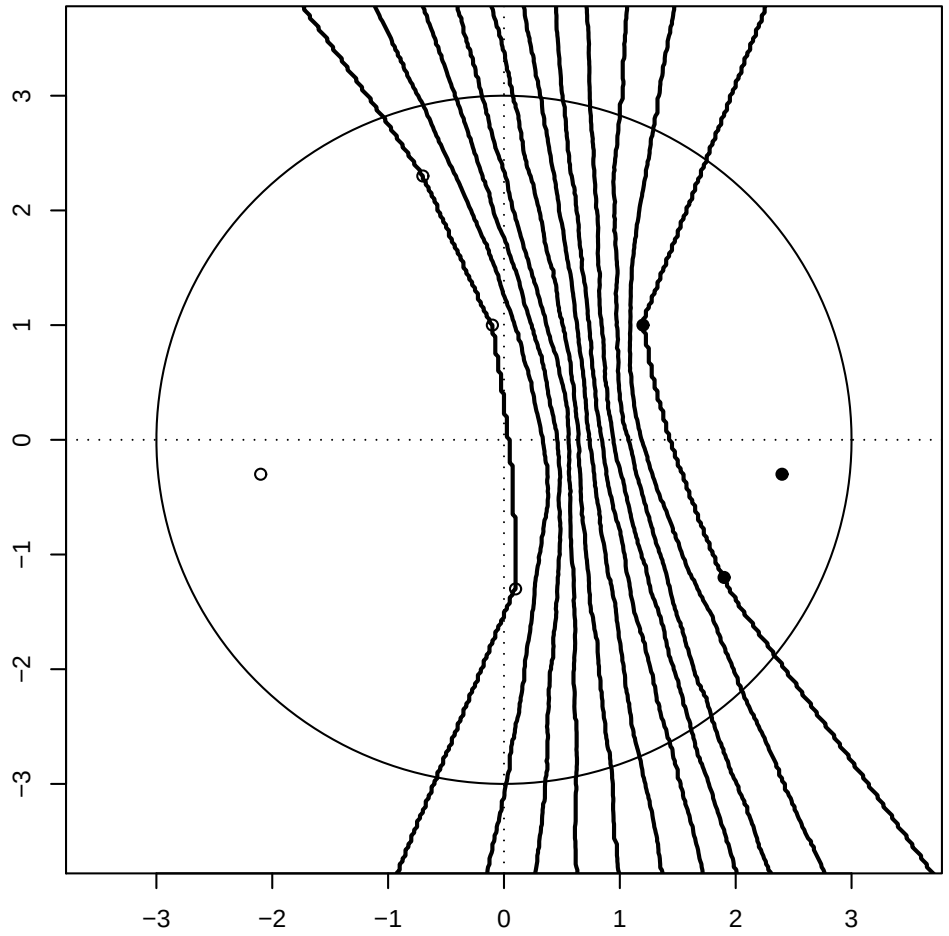
Using a sample of K values from the posterior, $(u^{(1)}, w^{(1)}), \ldots, (u^{(K)}, w^{(K)})$, we can approximate this as follows:

$$P(y^* = +1 \,|\, x^*, (x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})) \;\approx\; \frac{1}{K} \sum_{j=1}^{K} P(y^* = +1 \,|\, x^*, u^{(j)}, w^{(j)})$$

For this model, $P(y^* = +1 \,|\, x^*, u^{(j)}, w^{(j)})$ is either 0 or 1, depending on the sign of $u^{(j)}\,({w^{(j)}}^T x^* - 1)$. The average above is just the fraction of lines drawn from the posterior that would put the test point in class $+1$.

A Plot of the Predictive Probabilities

Here is a contour plot over the input space of the approximate predictive probability of class +1, based on a sample of size 10000 from the prior, which resulted in a sample of size 450 from the posterior:



The contour lines go from 0 on the left to 1 on the right, in steps of 0.1.

The Marginal Likelihood

The sample of 10000 values from the prior also lets us estimate the marginal likelihood for this model, given the seven observed data points.

We consider the $x^{(i)}$ to be fixed (not random), so the marginal likelihood is just the probability of all the $y^{(i)}$ having their observed values. This probability is one for a line that classifies all the points correctly, and zero for any other line.

We can therefore estimate the marginal likelihood by the fraction of lines drawn from the prior that are compatible with the data: $450/10000 = 0.045$.
[Actually... The marginal likelihood is half that, since the chance of picking u to have the +1's on the right side of the line is 1/2.]

We could use this to compare this model with some other, such as a model that said the classes were separated by quadratic rather than linear curves.

However... the marginal likelihood is **very sensitive** to the prior used. If we used a prior for the separating line that was uniform over a bigger region, say allowing the closest point to the origin to be up to a distance of 10 away, the marginal likelihood would be smaller. Computing marginal likelihoods makes sense only if you have given careful thought to the prior.

Final Thoughts on This Example

- We see that correctly translating informal knowledge into a prior distribution isn't always trivial.
- However, a prior can be *tested*, by checking how well it corresponds to our prior beliefs. Prior distributions are **not** "arbitrary".
- More elaborate priors might sometimes be appropriate. For example, we might use a prior that favoured lines that are almost horizontal or almost vertical, if we believe that probably one of the two inputs is mostly irrelevant.
- For a data set with seven points, only about 4.5% of the parameter values we drew from the prior made it into the posterior sample. This technique isn't going to work for realistic problems. We need better ways of sampling from the posterior distribution.

Comparison of Bayesian Learning with Other Approaches

Distinctive Features of the Bayesian Approach

Probability is used not only to describe "physical" randomness, such as errors in labeling, but also uncertainty regarding the true values of the parameters. These prior and posterior probabilities represent **degrees of belief**, before and after seeing the data.

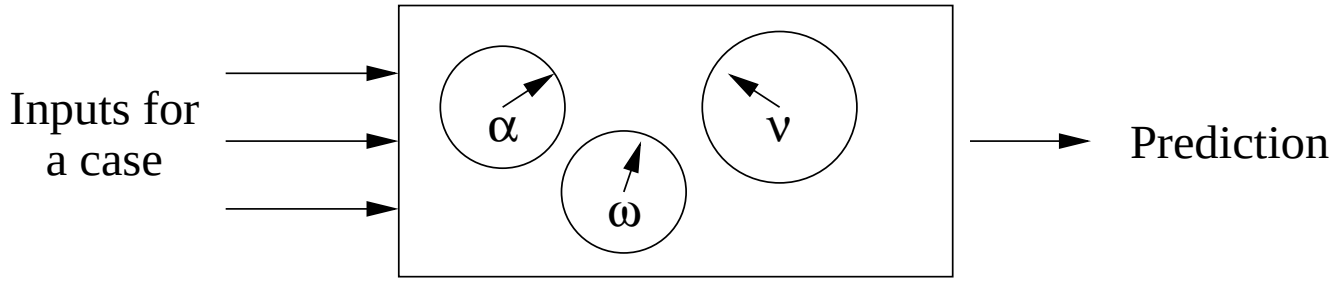
The Bayesian approach takes **modeling** seriously. A Bayesian model includes a suitable prior distribution for model parameters. If the model/prior are chosen without regard for the actual situation, there is *no justification* for believing the results of Bayesian inference.

The model and prior are chosen based on our knowledge of the problem. These choices are **not**, in theory, affected by the amount of data collected, or by the question we are interested in answering. We **do not**, for example, restrict the complexity of the model just because we have only a small amount of data.

Pragmatic compromises are inevitable in practice — eg, no model and prior perfectly express to our knowledge of the situation. The Bayesian approach relies on reducing such flaws to a level where we think they won't seriously affect the results. If this isn't possible, it may be better to use some other approach.

Contrast With the "Learning Machine" Approach

One view of machine learning pictures a "learning machine", which takes in inputs for a training/test case at one end, and outputs a prediction at the other:



The machine has various "knobs", whose settings change how a prediction is made from the inputs. Learning is seen as a procedure for twiddling the knobs in the hopes of making better predictions on test cases — for instance, we might use the knob settings that minimize prediction error on training cases.

This approach differs profoundly from the Bayesian view:

- The choice of learning machine is essentially *arbitrary* — unlike a model, the machine has no meaningful semantics, that we could compare with our beliefs.
- The "knobs" on the machine *do not* correspond to the parameters of a Bayesian model — Bayesian predictions, found by averaging, usually cannot be reproduced using *any* single value of the model parameters.

Contrast With "Learning Theory"

An aim of "learning theory" is to prove that certain learning machines "generalize" well. One can sometimes prove that if you adjust the knobs on the learning machine to minimize training error, then apply it to test cases, the training error rates and test error rates are unlikely to be far apart:

$$P(|\text{test error rate} - \text{training error rate}| > \epsilon) \quad < \quad \delta$$

where δ and ϵ have certain small values, which depend on the training set size.

Such a result would be of little interest, if it weren't usually interpreted as guaranteeing that, for instance:

$$P(|\text{test error rate} - 0.02| > \epsilon \mid \text{training error rate} = 0.02) \quad < \quad \delta$$

This is a fallacy, however — no valid probabilistic statement about test error rates conditional on the observed error rate on training cases is possible without assuming some prior distribution over possible situations. This fallacy is analogous to the common misinterpretation of a frequentist p-value as the probability that the null hypothesis is true, or of a frequentist confidence interval as an interval that likely contains the true value.

What About "Bias" and "Variance"?

Another approach to analysing learning methods (especially for predicting real-valued quantities) looks at the following two indicators of how well a method predicts some quantity:

Bias: how much predictions depart from the truth on average.

Variance: the average squared difference of predictions from their average.

The average squared error for the method can be decomposed as the sum of the squared bias and the variance. This leads to a strategy: choose a method that minimizes this sum, possibly trading off increased bias for reduced variance, or vice versa, by adjusting complexity, or introducing some form of "regularization".

There are two problems with this strategy:

- The bias and variance depend on the true situation, which is unknown.
- There is no reason to think that trying nevertheless to minimize squared bias plus variance produces a unique answer.

Assessments of bias and variance play no role in the Bayesian approach.

Some Issues Applying Bayesian Methods

The Challenge of Specifying Models and Priors

The first challenge in making the Bayesian approach work is to choose a suitable model and prior. This can be especially difficult for the complex, high-dimensional problems that are traditional in machine learning.

> **A suitable model** should encompass all the possibilities that are thought to be at all likely. Unrealistically limited forms of functions (eg, linear) or distributions (eg, normal) should be avoided.
>
> **A suitable prior** should avoid giving zero or tiny probability to real possibilities, but should also avoid spreading out the probability over all possibilities, however unrealistic. To avoid a prior that is too spread out, dependencies between parameters may need to be modeled.

Unfortunately, the effort in doing a good job can easily get out of hand. One strategy is to introduce *latent variables* into the model, and *hyperparameters* into the prior. Both of these are devices for modeling dependencies in a tractable way.

In practice, an iterative approach may be needed — we formulate our model and prior, try it out on data, and see by examining *diagnostics* that we've made a mistake. We go back and revise our model or prior, trying to avoid having our choice be unduly influenced by the data (since the data would then count twice).

Infinite Models

Many real phenomena are of essentially unlimited complexity:

> Suppose we model consumer behaviour by categorizing consumers into various "types" (mixture components). There is no reason to think that there are only (say) five types of consumer. Surely there are an unlimited number of types, though some may be rare.
>
> Suppose we model the growth rate of trees as a function of climate, soil type, genetic characteristics, disease suppression measures taken, etc. There is no reason to think any simple functional form (eg, linear, low-order polynomial) will capture the many ways these factors interact to determine tree growth.

How can we build a model that accommodates such complexity? One approach:

- Define models that can have any finite amount of complexity (eg, a finite number of mixture components, or of hidden units).
- Define priors for these models that make sense.
- See if the limit as the complexity goes to infinity is sensible.

If the limit makes sense, we can use a model that is as large as we can handle computationally. And sometimes, we can figure out how to actually implement the infinite model on a finite computer!

The Computational Challenge

The other big challenge in making Bayesian modeling work is computing the posterior distribution. There are four main approaches:

Analytical integration: Works when "conjugate" prior distributions can be used, which combine nicely with the likelihood — usually too much to hope for.

Gaussian approximation: Works well when there's a lot of data, compared to the model complexity — the posterior distribution in then close to Gaussian, and can be handled by finding its mode, and the second derivatives at the mode.

Monte Carlo integration: Once we have a sample of parameter values from the posterior distribution, most things are possible. But how to get a sample?

- *Simple Monte Carlo* — sample directly from the posterior. Seldom possible.
- *Importance sampling* — sample from an approximation to the posterior, then reweight to compensate for the difference. Maybe OK in moderate dimensions.
- *Markov chain Monte Carlo (MCMC)* — simulate a Markov chain that eventually converges to the posterior distribution. Can be applied to a remarkable variety of problems. Currently the dominant approach.

Variational approximation: A cleverer way to approximate the posterior. May sometimes be faster than MCMC, but it's not as general, and not exact.

Analytically-Tractable Bayesian Models

Conjugate Prior Distributions

For most Bayesian inference problems, the integrals needed to do inference and prediction are not analytically tractable — hence the need for numerical quadrature, Monte Carlo methods, or various approximations.

Most of the exceptions involve *conjugate priors*, which combine nicely with the likelihood to give a posterior distribution of the same form. Examples:

1) Independent observations from a finite set, with Beta / Dirichlet priors.

2) Independent observations of Gaussian variables with Gaussian prior for the mean, and either known variance or inverse-Gamma prior for the variance.

3) Linear regression with Gaussian prior for the regression coefficients, and Gaussian noise, with known variance or inverse-Gamma prior for the variance.

It's nice when a tractable model and prior are appropriate for the problem. Unfortunatley, people are tempted to use such models and priors even when they aren't appropriate.

Independent Binary Observations with Beta Prior

We observe binary (0/1) variables Y_1, Y_2, ..., Y_n.

We model these as being *independent*, and *identically distributed*, with

$$P(Y_i = y \,|\, \theta) \;\; = \;\; \left\{ \begin{array}{ll} \theta & \text{if } y = 1 \\ 1 - \theta & \text{if } y = 0 \end{array} \right\} \;\; = \;\; \theta^y \, (1-\theta)^{1-y}$$

Let's suppose that our prior distribution for θ is Beta(a,b), with a and b being known postive reals. With this prior, the probability density over $(0, 1)$ of θ is:

$$P(\theta) \;\; = \;\; \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \, \theta^{a-1} (1-\theta)^{b-1}$$

Here, the Gamma function, $\Gamma(c)$, is defined to be $\int_0^\infty x^{c-1} \exp(-x) \, dx$. For integer c, $\Gamma(c) = (c-1)!$.

Note that when $a = b = 1$ the prior is uniform over $(0, 1)$.

The prior mean of θ is $a/(a+b)$. Big a and b give smaller prior variance.

Posterior Distribution with Beta Prior

With this Beta prior, the posterior distribution is also Beta:

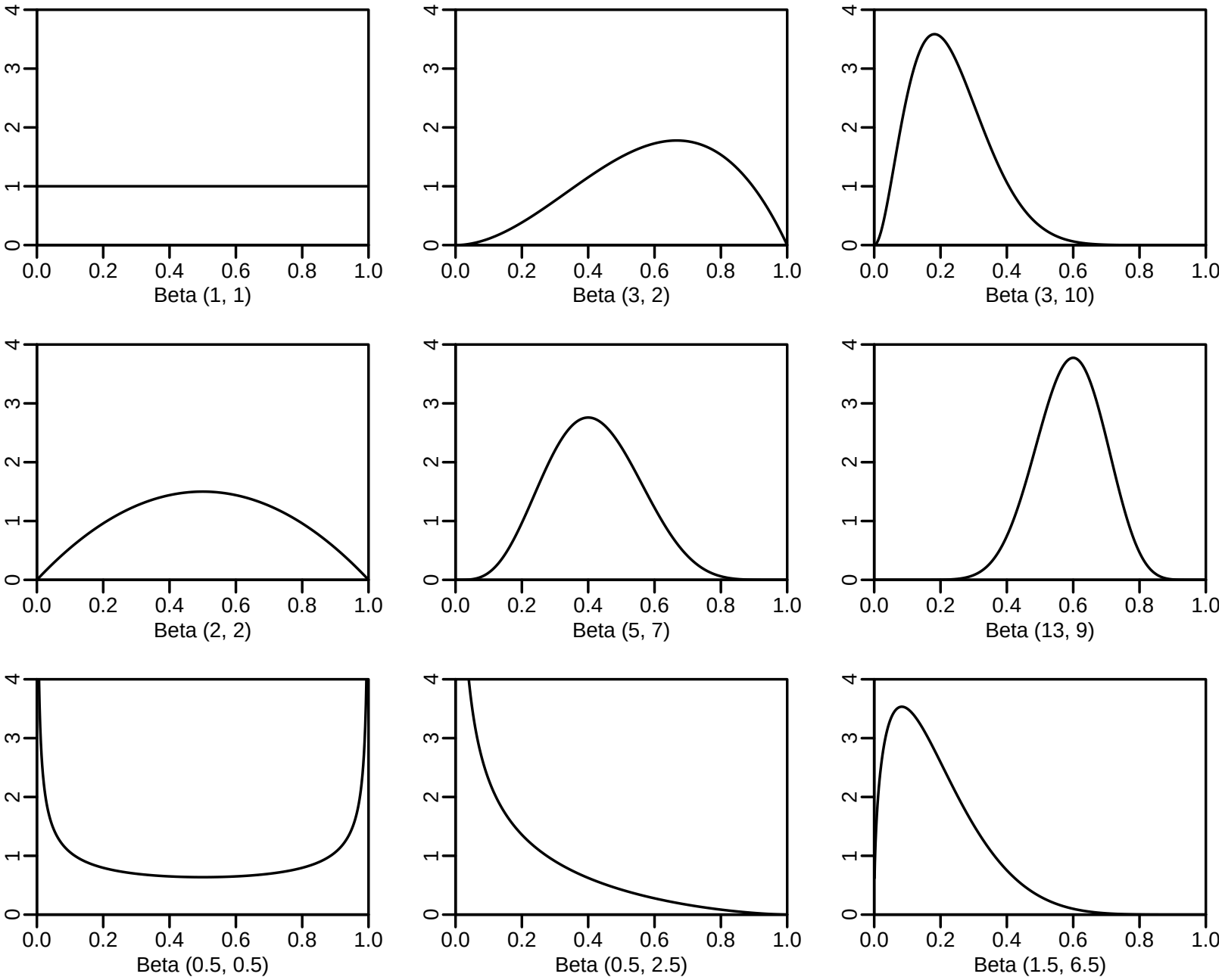
$$\begin{aligned} P(\theta \,|\, Y_1 = y_1,\, Y_2 = y_2,\, \ldots,\, Y_n = y_n) \\ \propto \quad & P(\theta) \prod_{i=1}^{n} P(Y_i = y_i \,|\, \theta) \\ \propto \quad & \theta^{a-1}\,(1-\theta)^{b-1} \prod_{i=1}^{n} \theta^{y_i}\,(1-\theta)^{1-y_i} \\ \propto \quad & \theta^{\Sigma y_i + a - 1}\,(1-\theta)^{n - \Sigma y_i + b - 1} \end{aligned}$$

So the posterior distribution is Beta $(\sum y_i + a,\, n - \sum y_i + b)$.

One way this is sometimes visualized is as the prior being equivalent to a fictitious observations with $Y = 1$ and b fictitious observations with $Y = 0$.

Note that all that is used from the data is $\sum y_i$, which is a *minimal sufficient statistic*, whose values are in one-to-one correspondence with possible likelihood functions (ignoring constant factors).

Examples of Beta Priors and Posteriors



Predictive Distribution from Beta Posterior

From the Beta $(\sum y_i + a,\, n - \sum y_i + b)$ posterior distribution, we can make a probabilitistic prediction for the next observation:

$$
\begin{aligned}
& P(Y_{n+1} = 1 \,|\, Y_1 = y_1,\, Y_2 = y_2,\, \ldots,\, Y_n = y_n) \\
&\quad = \int_0^1 P(Y_{n+1} = 1 \,|\, \theta)\, P(\theta \,|\, Y_1 = y_1,\, Y_2 = y_2,\, \ldots,\, Y_n = y_n)\, d\theta \\
&\quad = \int_0^1 \theta\, P(\theta \,|\, Y_1 = y_1,\, Y_2 = y_2,\, \ldots,\, Y_n = y_n)\, d\theta \\
&\quad = \int_0^1 \theta\, \frac{\Gamma(n+a+b)}{\Gamma(\Sigma y_i + a)\Gamma(n - \Sigma y_i + b)}\, \theta^{\Sigma y_i + a - 1}\, (1-\theta)^{n - \Sigma y_i + b - 1}\, d\theta \\
&\quad = \frac{\Gamma(n+a+b)}{\Gamma(\Sigma y_i + a)\Gamma(n - \Sigma y_i + b)}\, \frac{\Gamma(1 + \Sigma y_i + a)\Gamma(n - \Sigma y_i + b)}{\Gamma(1 + n + a + b)} \\
&\quad = \frac{\sum y_i + a}{n + a + b}
\end{aligned}
$$

This uses the fact that $c\Gamma(c) = \Gamma(1 + c)$.

Generalizing to More Than Two Values

For i.i.d. observations with a finite number, K, of possible values, with $K > 2$, the conjugate prior for the probabilities $\rho_1, \ldots, \rho_K$ is the Dirichlet distribution, with the following density on the simplex where all $\rho_k > 0$ and $\sum \rho_k = 1$:

$$P(\rho_1, \ldots, \rho_K) \quad = \quad \frac{\Gamma(\Sigma_k a_k)}{\Pi_k \Gamma(a_k)} \prod_{k=1}^{K} \rho_k^{a_k - 1}$$

The parameters $a_1, \ldots, a_K$ can be any positive reals.

The posterior distribution after observing n items, with n_1 having value 1, n_2 having value 2, etc. is Dirichlet with parameters $a_1 + n_1, \ldots, a_K + n_K$.

The predictive distribution for item $n + 1$ is

$$P(Y_{n+1} = k \,|\, Y_1 = y_1, \ldots, Y_K = y_k) \quad = \quad \frac{n_k + a_k}{n + \Sigma a_k}$$

Independent Observations from a Gaussian Distribution

We observe real variables Y_1, Y_2, ..., Y_n.

We model these as being independent, all from some Gaussian distribution with unknown mean, μ, and known variance, σ^2.

The conjugate prior for μ is Gaussian with some mean μ_0 and variance σ_0^2.

Rather than talk about the variance, it is more convenient to talk about the *precision*, equal to the reciprocal of the variance. A data point has precision $\tau = 1/\sigma^2$ and the prior has precision $\tau_0 = 1/\sigma_0^2$.

The posterior distribution for μ is also Gaussian, with precision $\tau_n = \tau_0 + n\tau$, and with mean

$$\mu_n = \frac{\tau_0 \mu_0 + n\tau \overline{y}}{\tau_0 + n\tau}$$

where $\overline{y}$ is the sample mean of the observatons $y_1, \ldots, y_n$.

The predictive distribution for Y_{n+1} is Gaussian with mean μ_n and variance $(1/\tau_n) + \sigma^2$.

Gaussian with Unknown Variance

What if both the mean and the variance (precision) of the Gaussian distribution for $Y_1, \ldots, Y_n$ are unknown?

There is still a conjugate prior, but in it, μ and τ are dependent:

$$\begin{aligned} \tau &\sim \text{Gamma}(a, b) \\ \mu \,|\, \tau &\sim N(\mu_0, c/\tau) \end{aligned}$$

for some constants a, b, and c.

It's hard to imagine circumstances where our prior information about μ and τ would have a dependence of this sort. But unfortunately, people use this conjugate prior anyway, because it's convenient.