T-61.5040

Learning Models and Methods, (5 cr) L

Spring 2007

Lecturer: **Petteri Pajunen**

Course Assistant: **Ville Viitaniemi**

LECTURE 6: 21.2.2007

BAYESIAN MODELING: ONE-VARIABLE MODELS, PRIORS

One-Variable Normal Models

- we will go through a few examples using Normal distribution illustrating certain aspects of Bayesian Inference
- benefits: closed-form solutions, easy to interpret what happens, can be used as building blocks in more complicated models
- drawbacks: Bayesian inference in practice almost always requires more complicated techniques

- example: Normally distributed data with known variance
- a Normal distribution $p(y|\theta, \sigma^2) = N(y|\theta, \sigma^2)$ is defined by its *mean* θ and *variance* σ^2
- we could interpret θ as the most probable value and variance as a measure of how certain we are about θ

- for convenience, we will use *precision* $\lambda = \sigma^{-2}$ instead of variance
- large precision $\implies$ small variance, θ known quite precisely
- small precision $\implies$ large variance, θ not known precisely
- lets examine what happens when we observe Normally distributed data

- likelihood is $p(y|\theta,\lambda) = N(y|\theta,\lambda^{-1})$ where

$$N(y|\theta,\lambda^{-1}) = \lambda^{1/2}(2\pi)^{-1/2}\exp(-\frac{\lambda}{2}(y-\theta)^2)$$

- assume that λ is known but θ is unknown and we observe one value y

- lets choose the prior as $p(\theta) = N(\theta|\theta_0,\lambda_0^{-1})$
- Bayes' Theorem gives the posterior as (exercise):

$$p(\theta|y) = N(\theta \mid \frac{\lambda_0\theta_0 + \lambda y}{\lambda_0 + \lambda}, (\lambda_0+\lambda)^{-1})$$

- note that both the prior and the posterior are Normal

- lets examine how the mean and precision change as we go from Normal prior to Normal posterior
- precision changes as

$$\lambda_0 \to \lambda_0 + \lambda$$

- λ is the "data precision": observing one y *increases* the precision by λ
- λ_0, the prior precision, defines how accurately we knew θ before seeing any data

- mean is changed towards y as

$$\theta_0 \to \theta_0 + (y - \theta_0)\frac{\lambda}{\lambda + \lambda_0}$$

- the posterior mean is a weighted average of θ_0 and y
- the weighting depends on the precisions

- the "step size"

$$\frac{\lambda}{\lambda + \lambda_0}$$

defines the weights

- if prior is more precise ($\lambda \ll \lambda_0$), the step size is close to zero and the posterior mean is close to prior mean
- if prior is not precise ($\lambda \gg \lambda_0$), the step size is close to one and the posterior mean is closer to y

- above discussion illustrates what happens in general
- compromise between data and the prior, automatically weighted by prior and likelihood precisions
- very precise prior makes data useless: we already know θ precisely
- very precise likelihood makes prior useless: data tells the value of θ very precisely

- what is the predictive distribution of a *new* value $\tilde{y}$?
- use "reverse marginalization" to compute

$$\begin{aligned} p(\tilde{y}|y) &= \int p(\tilde{y}, \theta|y)d\theta \\ &= \int p(\tilde{y}|y, \theta)p(\theta|y)d\theta \end{aligned}$$

- integrand contains $p(\tilde{y}|y, \theta) = N(\tilde{y}|\theta, \lambda^{-1})$

- the second term is the Normal posterior $p(\theta|y)$ which was just solved
- integrand is a Normal distribution of θ and $\tilde{y}$
- this can be seen by multiplying the two Normal distributions and writing the product in the form of $N((\theta, \tilde{y})'|A, B)$

- integration over θ gives a marginal distribution of the joint Normal distribution
- therefore $p(\tilde{y}|y)$ is also a Normal distribution
- its mean and variance are (see model averaging exercise problem)

$$\mathrm{E}(\tilde{y}|y) = \mathrm{E}(\theta|y)$$
$$\mathrm{var}(\tilde{y}|y) = \sigma^2 + \mathrm{var}(\theta|y)$$

- the predictive mean is identical to the posterior mean
- this corresponds to intuition: posterior mean is the most probable θ, and most probable $\tilde{y}$ generated is the mean of the Normal distribution
- but predictive variance is *not the likelihood variance*
- $\mathrm{var}(\tilde{y}|y)$ is larger than σ^2, since there is also uncertainty about θ

- the same calculations may be repeated for n independent observations $y_1, \ldots, y_n$ which can also be vectors (exercise)
- posterior precision will be a sum of prior precision and one data precision for each observation, i.e. $\lambda_0 + n\lambda$
- the more data, the higher the precision of the posterior
- demo `bgauss.R` illustrates the above inference

- lets use previous results to construct a simple Bayesian classifier
- warning: this example is simplified too much to be realistic
- the purpose is to build the classifier using the Normal model as a building block

- data x_i belongs to one of two classes (0 and 1)
- data in each class is Normally distributed:

$$x_i \sim N(\mu_j, \lambda_j^{-1}) \text{ if } x_i \text{ is in class } j = 0, 1$$

- knowns: precisions λ_0, λ_1, training data $(x_i, y_i),\ i = 1, \ldots, n$
- unknowns: μ_0, μ_1 and class label $\tilde{y} \in \{0, 1\}$ corresponding to a new $\tilde{x}$
- prior: $p(\mu_0, \mu_1) = N(\mu_0|0, 1)N(\mu_1|1, 1)$

- unrealistic assumption: assume that each class generates equal amount of data
- this means that any $p(y = 0| \ldots)$ *not conditional to* x is 0.5
- in a realistic classifier, this probability would be an unknown parameter

- the likelihood is

$$\begin{aligned} p(D|\theta) &= \prod_i p((x_i, y_i)|\mu_0, \mu_1) = \\ &= \prod_i p(x_i|y_i, \mu_0, \mu_1)p(y_i|\mu_0, \mu_1) \end{aligned}$$

- according to our assumptions

$$\begin{aligned} p(x_i|y_i = 0, \mu_0, \mu_1) &= N(x_i|\mu_0, \lambda_0^{-1}) \\ p(x_i|y_i = 1, \mu_0, \mu_1) &= N(x_i|\mu_1, \lambda_1^{-1}) \\ p(y_i|\mu_0, \mu_1) &= 0.5 \end{aligned}$$

- then the posterior can be written as

$$p(\theta|D) \propto \left[\prod_i p((x_i, y_i)|\mu_0, \mu_1)\right] p(\mu_0, \mu_1) =$$
$$\propto \left[\prod_i N(x_i|\mu_{y_i}, \lambda_{y_i}^{-1})\right] N(\mu_0|0,1)N(\mu_1|1,1)$$

- the posterior splits into factors $p(\mu_0|D)p(\mu_1|D)$
- both can be computed using parameters and data corresponding to one class only
- define n_j as number of vectors in class j, and s_j as sum of vectors in class j

- now each posterior factor is simply the result of observing Normal data and having a Normal prior
- using the formulas derived earlier, we obtain

$$p(\mu_0|D) = N(\mu_0|\frac{\lambda_0 s_0}{n_0\lambda_0 + 1}, (1 + n_0\lambda_0)^{-1}) = N(\mu_0|m_0, v_0)$$
$$p(\mu_1|D) = N(\mu_1|\frac{\lambda_1 s_1 + 1}{n_1\lambda_1 + 1}, (1 + n_1\lambda_1)^{-1}) = N(\mu_1|m_1, v_1)$$

- to use the classifier, we need to compute the predictive distribution of $\tilde{y}$ conditional to the vector $\tilde{x}$ to be classified:

$$p(\tilde{y}|\tilde{x}, D)$$

- since either $\tilde{y} = 0$ or $\tilde{y} = 1$, it is enough to compute

$$p(\tilde{y} = 1|\tilde{x}, D) = \int p(\tilde{y} = 1|\tilde{x}, \mu_1)p(\mu_1|D)d\mu_1$$
$$\propto \int p(\tilde{x}|\tilde{y} = 1, \mu_1)p(\tilde{y} = 1|\mu_1)N(\mu_1|m_1, v_1)d\mu_1$$
$$\propto \int N(\tilde{x}|\mu_1, \lambda_1^{-1})N(\mu_1|m_1, v_1)d\mu_1$$

- as a function of $\tilde{x}$, this is a Normal distribution $N(\tilde{x}|m_1, v_1 + \lambda_1^{-1})$
- therefore we get

$$p(\tilde{y} = 1|\tilde{x}, D) \propto N(\tilde{x}|m_1, v_1 + \lambda_1^{-1})$$
$$p(\tilde{y} = 0|\tilde{x}, D) \propto N(\tilde{x}|m_0, v_0 + \lambda_0^{-1})$$

- the classifier can be implemented by choosing the class j which maximizes $p(\tilde{y} = j|\tilde{x}, D)$

- a curiosity: the classifier works even if your training data comes from one class only!
- if $n_0 = 0$, then $m_0 = 0$, $v_0 = 1$, and

$$p(\tilde{y} = 0|\tilde{x}, D) \propto N(\tilde{x}|0, \lambda_0^{-1} + 1)$$

- all the above can be easily extended to multivariate data, using the corresponding multivariate formulas given in the exercises
- demo: `bclass2.R`

Priors

- Bayes requires specifying the likelihood $p(D|\theta)$ and the prior $p(\theta)$
- the likelihood seems easier, since it tells how data is generated as a function of θ
- prior can be more difficult:
  - often models contain parameters that have no easily understood meaning
  - for arbitrary priors and likelihoods, it may be impossible to compute the posterior in closed form

Conjugate Priors

- closed-form posteriors can be obtained by using *conjugate priors*
- when just the mean is unknown, Normal prior and likelihood lead to a Normal posterior
- assume that likelihood belongs to class $\mathcal{L}$ of distributions and the prior to a class $\mathcal{P}$
- if it follows that the posterior belongs to class $\mathcal{P}$, then we say that the classes $\mathcal{L}$ and $\mathcal{P}$ are conjugate

- if in the above definition $\mathcal{L} = \mathcal{P}$, then the prior is *naturally conjugate* to the likelihood
- e.g. Normal and binomial distributions are naturally conjugate
- conjugacy is practical since one can simply compute the posterior parameters (e.g. mean and variance)
- a realistic model usually results in priors that are not conjugate

- conjugacy often reveals the effect of the prior
- the prior can be interpreted as representing previously observed data
- for example in the Normal example, the posterior mean is symmetric with respect to the prior and the likelihood:
$$\mathrm{E}(\theta|y) = \frac{\lambda_0\theta_0 + \lambda y}{\lambda_0 + \lambda}$$
- prior contains the same information as having observed θ_0 with precision λ_0

Noninformative Priors

- Bayes is often criticized for its use of prior distribution
- by a suitable prior, the end result can be influenced
- but this is true of the whole model as well: the result does (and should) depend on the choice of model
- a better argument is that it is difficult to specify prior information for more or less abstract parameters when one has no idea whatsoever about the value

- there has been considerable interest to solve the question of determining *noninformative priors*
- it is difficult to even define what a noninformative prior is
- heuristic ideas don't work: for example, shouldn't all values be equally probable if we have no prior information?

- suppose the variance σ^2 is the unknown quantity of interest
- we might as well use the standard deviation σ
- use a constant prior, so any unit interval has the same probability
- since $\sigma^2 \in [0, 1]$ and $\sigma \in [0, 1]$ are the same events, their probability must be the same q

- constant prior says that $\sigma^2 \in [1,2]$ and $\sigma \in [1,2]$ have prior probability q as well
- but these events are not the same: $\sigma \in [1,2]$ is equivalent to $\sigma^2 \in [1,4]$
- the latter event has prior probability $3q \neq q$
- results should not depend on whether you want to use σ^2 or σ

- the *Fisher information* for a scalar θ is

$$I(\theta) = \mathrm{E}\left[\frac{\partial}{\partial\theta}\log p(y|\theta)\right]^2$$

  where the expectation is computed over $p(y|\theta)$
- the *Jeffreys' prior* $p(\theta) \propto \sqrt{I(\theta)}$ is invariant to transformations of parameters
- for example, the problem of σ vs σ^2 is solved by it

- there are many ways of defining noninformative priors, most of which disagree with Jeffreys' prior
- for scalar *location* and *scale* parameters, it seems that most ways lead to the same result
- the Jeffreys' prior for a location parameter is $p(\mu) \propto$ constant and for a scale parameter it is $p(\theta) \propto 1/\theta$ (exercise problem)

- example: table entry problem
- it has been found experimentally that the distribution of the first significant digit in tables of scale data is well described by

$$\log(1 + i^{-1}) / \log 10, \; i = 1, \ldots, 9$$

- one might expect a uniform distribution instead
- scale data: populations of cities, number of cars passing a bridge in a day etc..

- Jeffreys' prior for scale parameter is $p(\sigma) \propto \sigma^{-1}$
- consider the interval $(1, 10)$ where Jeffreys' prior is

$$p(\sigma) = \sigma^{-1} / \log 10$$

- the first digit of σ is i if $\sigma \in [i, i+1)$

- the probability of this is

$$\begin{aligned} p(\sigma \in [i, i+1)) &= \int_i^{i+1} \sigma^{-1} / \log 10 d\sigma \\ &= /_i^{i+1} \log \sigma / \log 10 \\ &= (\log i + 1 - \log i) / \log 10 \\ &= \log(1 + i^{-1}) / \log 10 \end{aligned}$$

- so Jeffreys' prior restricted to the interval $[1, 10)$ predicts the experimental result

Improper Priors

- sometimes one may encounter situations where the constant prior $p(\theta) = c$ seems the correct choice
- this cannot be normalized to a probability distribution if θ varies over an infinite interval (e.g. $\theta \in \mathbb{R}$)
- however, such priors are sometimes used: they are called *improper priors*

- in most cases they lead to proper posteriors: when they do not, the results may be wrong
- it may be better to choose a flat proper prior, and allow the likelihood to dominate the inference as it would for a constant prior
- example: a Normal prior with a very large variance may be close enough to a constant prior