Administrivia

HW three now posted

Computer vision seminar of possible interest on Friday
(Ernesto on integration in high dimensions)

In general, this seminar runs Fridays 2-3:30 in GS 906. If you want to receive emails about it, join compvision@list.arizona.edu (or ask me to add you).

Bayesian statistics summary

- Bayesian statistical models
  - We prefer generative models for likelihood (and prior)
  - Conjugate priors are preferred when they are accurate enough
  - Bayesian updating for sequences of independent data
    - Yesterday’s posterior becomes today’s prior
- Inference uses Bayes rule to “invert” the forward model
  - Result is the posterior distribution
  - MAP estimate provides a single “best” number (often not the best)

Bayesian statistics summary

- Related topics coming up
  - Predictive distribution
    - Marginalizes out uncertainty about models
  - Model selection
  - Estimation and decision making

Bayesian Sequential Update

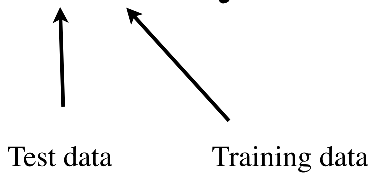
$$p(\theta \mid D_{1:N}) \propto \left\{ p(\theta) \prod_{i=1}^{N-1} \left(p(D_i \mid \theta) \right) \right\} p(D_N \mid \theta)$$

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

Posterior from 1:N-1 is now the prior

Predictive Distribution

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



Test data

Training data

Model Selection

- Model selection refers to choosing among different instances within a model class (1) or different model classes (2).

- Examples:
  - The number of clusters (1)
  - The degree of a polynomial to fit a curve to data (1)
  - Polynomials versus other basis functions such as Fourier (2)

Model Comparison Difficulties

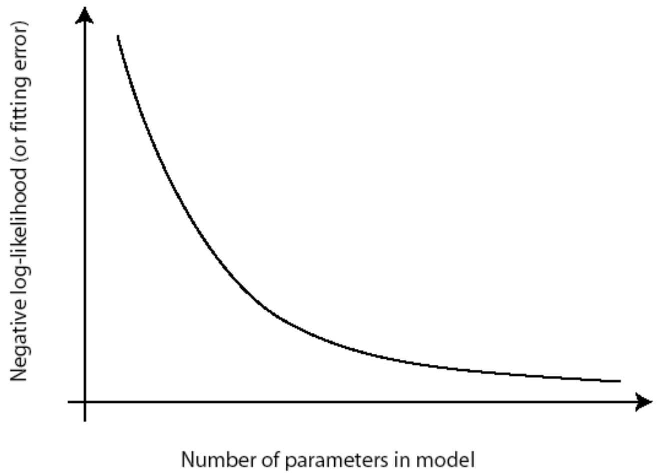
- Prior densities of different models are typically of different dimensionality (leads to expensive integration).

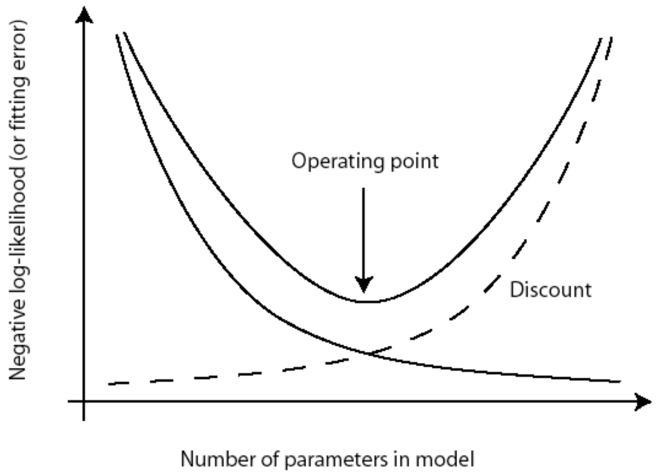
- Good likelihoods help select models, but constructing them is an exacting task.
  - Don’t forget about the “negative space”
    - A more complex model (e.g., more objects in a scene) explains more data, but it also proposes more data where there is none.
    - Missing data must be penalized!

- Good priors over different model classes are often not obvious

Solutions (penalize complexity)

- Typical approach is to focus on the balance between fitting accuracy, and model complexity using various penalties.





Solutions (penalize complexity)

- Typical approach is to focus on the balance between fitting accuracy, and model complexity using various penalties.
- AIC (An information criterion, Akaike, 74)

Replace log likelihood, $\log\left(\mathrm{p}(\mathrm{D}|\theta)\right)$, with $\log\left(\mathrm{p}(\mathrm{D}|\theta)\right) - M$
where M is the number of adjustable parameters.

Solutions (penalize complexity)

- Typical approach is to focus on the balance between fitting accuracy, and model complexity using various penalties.
- BIC (Bayesian information criterion)

Replace log likelihood, $\log\left(\mathrm{p}(\mathrm{D}|\theta)\right)$, with $\log\left(\mathrm{p}(\mathrm{D}|\theta)\right) - \frac{1}{2} M \log(N)$
where M is the number of adjustable parameters, N is the number of data points. This is the usual approximation. See Bishop, page 216-217 for a more complicated version.

Often also called minimum discription length (MDL)

The dependency on N may seem confusing. Note that the likelihood typically depends on N (often N is an exponent), but the formula above does not expose this.

Solutions (penalize complexity)

- Typical approach is to focus on the balance between fitting accuracy, and model complexity using various penalties.

- DIC (Deviation information criterion)
  - Details omitted (see Google)
  - Slightly more complex, but easier to compute using MCMC sampling
  - Still relies on strong assumptions (distribution is approximately multivariate normal)

Solutions (likelihood function)

- Incorrect complex models may predict lots of data where there is none
- Solution is to model missing data
- Example --- finding asteroids from detections amidst noise
  - Predicting more asteroids explains more data, but we expect to see detections for them most of the time.
  - Good modeling the probability of noise detections and probability of missing detections has a greater affect on the posterior than a prior (necessarily not very strong) on the number of asteroids.

Solutions (integrating parameter uncertainty)

$$p(D|M_i) = \int_{\Omega_i} p(D|\theta)\, p(\theta|M_i)\, d\theta \qquad \text{(Model evidence)}$$

and we can evaluate $\mathrm{p}(M_i|D)$ by Bayes.

The dimension of the space of θ (Ω_i in the integral) is typically a function of i.

This is argued (Bishop, §3.4) to be a principled way to penalize complex models because complex models spread their probability mass over greater support (but thec skeptic asks when or why the amount of penalty is correct).

Under additional approximations and assumptions, this becomes BIC (Bishop, §4.4.1).

Solutions (integrating parameter uncertainty)

$$p(D|M_i) = \int_{\Omega_i} p(D|\theta)\, p(\theta|M_i)\, d\theta \qquad \text{(Model evidence)}$$

and we can evaluate $\mathrm{p}(M_i|D)$ by Bayes.

We can compare two models abilities to explain data by the Bayes factor

$$K_{ij} = \frac{p(D|M_i)}{p(D|M_j)}$$

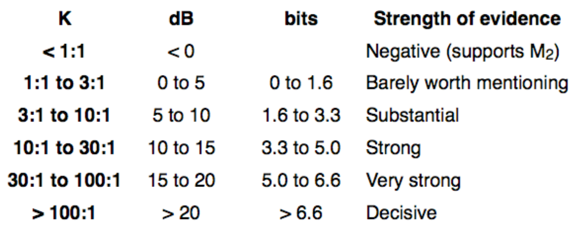
(We can augment with factors for the priors $p(M)$ if known)

Supplementary material on lecture notes page has a link to a classic reference on Bayes factors (Kass and Raftery, 95).

Solutions (integrating parameter uncertainty)

$$K_{ij} = \frac{p\left(D \middle| M_i\right)}{p\left(D \middle| M_j\right)} \qquad \text{(Bayes factor)}$$

Rules of thumb for K (from Jeffreys, via WikiPedia)



K	dB	bits	Strength of evidence
< 1:1	< 0		Negative (supports M2)
1:1 to 3:1	0 to 5	0 to 1.6	Barely worth mentioning
3:1 to 10:1	5 to 10	1.6 to 3.3	Substantial
10:1 to 30:1	10 to 15	3.3 to 5.0	Strong
30:1 to 100:1	15 to 20	5.0 to 6.6	Very strong
> 100:1	> 20	> 6.6	Decisive

Solutions (model averaging)

Recall the predictive distributions

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

To mitigate uncertainty of different models

$$p\left(x \middle| X\right) = \sum_i p\left(M_i\right) \int_{\Omega_i} p(x \mid \theta_i) p(\theta_i \mid X, M_i) d\theta$$

Note the assumption that M_i influences x through θ_i only, so no conditioning on M_i in the first factor in the integral.

Comments on Bayes factors, etc.

- Bayes factors can be used to derive BIC under specific conditions

- Otherwise you will normally need a numeric approximation of the integral

- $\mathrm{p}\left(D \middle| M_i\right)$ tells you the probability of observing the data you did under a well specified, and possibly flawed model—it is hard to know you compared the right alternatives.

- $\mathrm{p}\left(D \middle| M_i\right)$ does not necessarily tell you how well the model will predict other data

Cross-validation

- Standard way to evaluate models
- Exclude a subset of the data while fitting model
- Compute predictions for the held-out subset.
- Evaluate predictions against actual held-out values
  - e.g., distance from truth, or class labels
- If you use k such sets, this is called k-fold cross-validation
- If you leave out 1 data point, it is called leave-one-out.

Cross-validation (2)

- Cross-validation provides
  - A way to choose models
  - A way to measure performance
  - A way to measure generalization capacity
- Held out data **must be different enough** to test the level of generality that you want

Cross-validation (2)

- Cross-validation provides
  - A way to choose models
  - A way to measure performance
  - A way to measure generalization capacity
- Held out data **must be different enough** to test the level of generality that you want

  - Consider degree of validation in a model to predict happiness
    1. How happy are you now given recent data points
    2. How happy are you now given all data points
    3. How happy are you on day X given data for other days
    4. How happy are you based on model of **other** people
    5. How happy are you based on **other** people in other experiments
    6. How happy are you based on modeling people in other cultures