STSCI 4780/5780: Composite hypotheses, cont'd: Model comparison and prediction

Tom Loredo, CCAPS & SDS, Cornell University

Recap: Inference with parametric models

Single-model inference

Context = choice of single model (specific i)
Parameter estimation: What can we say about θ_i or $f(\theta_i)$?
Prediction: What can we say about future data D'?

Multi-model inference ("M-closed")

Context = $M_1 \vee M_2 \vee \cdots$
Model comparison/choice: What can we say about i?
Model averaging:

- *Systematic error*: $\theta_i = \{\phi, \eta_i\}$; ϕ is common to all
  What can we say about ϕ w/o committing to one model?
- *Prediction*: What can we say about future D', accounting for model uncertainty?

Model checking ("M-open")

Premise = $M_1 \vee$ "all" alternatives
Is M_1 adequate? (predictive tests, calibration, robustness)

Recap: Simple vs. composite hypotheses

Simple hypotheses

For a set of simple hypotheses, specifying the hypothesis completely determines the sampling distribution (conditional predictive distribution) for possible data: $P(D|H_i)$ can be directly evaluated when i is specified

Composite/compound hypotheses

Specifying a *composite* hypothesis narrows down the choice of the sampling distribution or likelihood function, but requires further information for the distribution to be fully determined

LTP and composite hypotheses

We can resolve a composite hypothesis into simple components, using LTP to compute it's overall probability. For a parametric model composite hypothesis, H, we often use

$$P(H|D,\mathcal{C}) = \int \mathrm{d}\theta\, p(H,\theta|D,\mathcal{C}) = \int \mathrm{d}\theta\, p(\theta|D,\mathcal{C})\, p(H|\theta,D,\mathcal{C})$$

Agenda: Composite hypotheses and uncertainty propagation

- *Lec08*: Marginalizing over nuisance parameters
- *Today*:
  - Model comparison
  - Model averaging (briefly!)
  - Prediction

Model comparison

Problem statement

$\mathcal{C} = (M_1 \vee M_2 \vee \ldots)$ — Specify a set of models.
$H_i = M_i$ — Hypothesis chooses a model.

Posterior probability for a model

$$p(M_i|D,\mathcal{C}) = p(M_i|\mathcal{C})\frac{p(D|M_i,\mathcal{C})}{p(D|\mathcal{C})}$$
$$\propto p(M_i|\mathcal{C})\mathcal{L}(M_i)$$

$$\mathcal{L}(M_i) \equiv p(D|M_i) = \int d\theta_i \, p(\theta_i|M_i)p(D|\theta_i, M_i)$$

Likelihood for model = Average likelihood for its parameters

$$\mathcal{L}(M_i) = \langle \mathcal{L}(\theta_i) \rangle$$

Varied terminology: Prior predictive = *Marginal likelihood* = Average likelihood = Global likelihood = (Weight of) Evidence for model

Odds and Bayes factors

A ratio of probabilities for two propositions using the same premises is called the *odds* favoring one over the other:

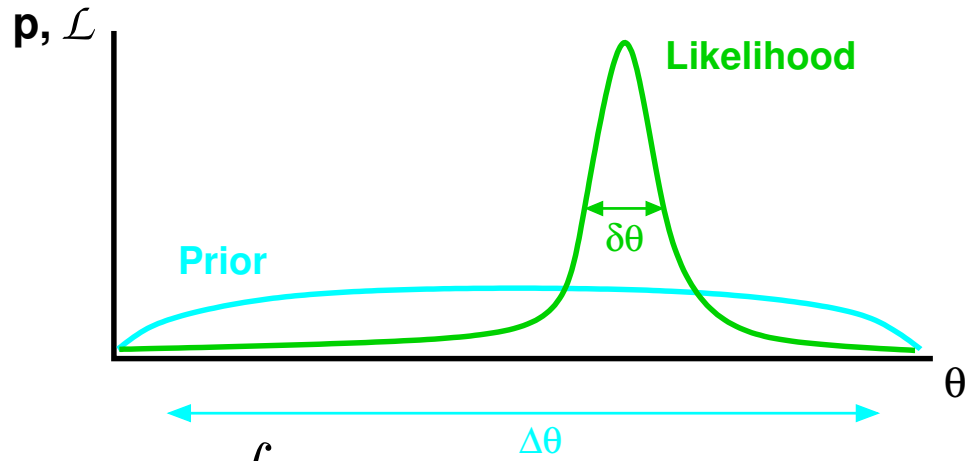
$$\begin{aligned} O_{ij} &\equiv \frac{p(M_i|D,\mathcal{C})}{p(M_j|D,\mathcal{C})} \\ &= \frac{p(M_i|\mathcal{C})}{p(M_j|\mathcal{C})} \times \frac{p(D|M_i,\mathcal{C})}{p(D|M_j,\mathcal{C})} \end{aligned}$$

The data-dependent part is called the *Bayes factor*:

$$B_{ij} \equiv \frac{p(D|M_i,\mathcal{C})}{p(D|M_j,\mathcal{C})}$$

It is a *likelihood ratio*; the BF terminology is usually reserved for cases when the likelihoods are marginal/average likelihoods for *composite hypotheses*

The Ockham Factor



$$
\begin{aligned}
p(D|M_i) &= \int d\theta_i \; p(\theta_i|M_i)\; \mathcal{L}(\theta_i) \approx p(\hat{\theta}_i|M_i)\mathcal{L}(\hat{\theta}_i)\delta\theta_i \\
&\approx \mathcal{L}(\hat{\theta}_i)\frac{\delta\theta_i}{\Delta\theta_i} \\
&= \text{Maximum Likelihood} \times \text{Ockham Factor}
\end{aligned}
$$

Models with more parameters often make the data more probable — *for the best fit*

Ockham factor penalizes models for "wasted" *volume of parameter space*

Quantifies intuition that models shouldn't require fine-tuning

Example: Equal probabilities for binary outcomes?

M_1: $\alpha = 1/2$ (a simple hypothesis)
M_2: $\alpha \in [0, 1]$ with flat prior
$\mathcal{C}$: $M_1 \vee M_2$; $D =$ FFSSSSFSSSFS — 8 successes in 12 trials

Maximum Likelihood ratio

From Bernoulli trials model:

$$M_1 : \qquad p(D|M_1) = \frac{1}{2^N} = 2.44 \times 10^{-4}$$

$$M_2 : \quad \mathcal{L}(\hat{\alpha}) \;=\; \left(\frac{2}{3}\right)^n \left(\frac{1}{3}\right)^{N-n} = 4.82 \times 10^{-4}$$

$$\frac{p(D|M_1)}{p(D|\hat{\alpha}, M_2)} = 0.51$$

Maximum likelihood (slightly) favors M_2 (on the basis of best-fit α)

Binary outcomes Bayes factor

$$p(D|M_1) = \frac{1}{2^N}; \qquad \text{and} \qquad p(D|M_2) = \frac{n!(N-n)!}{(N+1)!}$$

$$\begin{aligned} \rightarrow B_{12} \equiv \frac{p(D|M_1)}{p(D|M_2)} &= \frac{(N+1)!}{n!(N-n)!2^N} \\ &= 1.57 \end{aligned}$$

Bayes factor (odds) favors M_1 (equiprobable)

Note that for $n = 6$, $B_{12} = 2.93$; for this small amount of data, we can never be very sure results are equiprobable

If $n = 0$, $B_{12} \approx 1/315$; if $n = 2$, $B_{12} \approx 1/4.8$; for extreme data, 12 flips *can* be enough to lead us to strongly suspect outcomes have different probabilities

Example: Signal detection with Gaussian noise

Data consist of N measurements with additive noise:

$$d_i = \mu + \epsilon_i, \qquad i = 1 \text{ to } N$$

Noise contributions are independent, and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, with σ known

Consider two models:

- M_1: $\mu = \mu_1$ (perhaps zero)
- M_2: μ is uncertain with flat prior over $[\mu_l, \mu_u]$ (search range)

Likelihood functions

The form of the sampling dist'n is the same for both models (they just say different things about what μ to use):

$$\begin{aligned} \mathcal{L}(\mu) &= \prod_i p(d_i|\mu, \sigma, \mathcal{C}) \\ &= \frac{1}{\sigma^N (2\pi)^{N/2}} \exp\left(-\frac{Nr^2}{2\sigma^2}\right) \exp\left(-\frac{N(\mu - \overline{d})^2}{2\sigma^2}\right) \end{aligned}$$

M_1 marginal likelihood

M_1 just sets $\mu = \mu_1$ (simple hypothesis); equivalently, it uses a δ function prior: $p(\mu|M_1) = \delta(\mu - \mu_1)$:

$$\mathcal{L}(M_1) = C \exp\left(-\frac{N(\mu_1 - \overline{d})^2}{2\sigma^2}\right)$$

M_2 marginal likelihood

M_2 has $p(\mu|M_2) = 1/\Delta$ inside $[\mu_l, \mu_u]$, with $\Delta \equiv \mu_u - \mu_l$, so

$$\begin{aligned}
\mathcal{L}(M_2) &\equiv p(D|M_2) = \int \mathrm{d}\mu \; p(\mu|M_2)\, \mathcal{L}(\mu) \\
&= \frac{C}{\Delta} \int_{\mu_l}^{\mu_u} \exp\left(-\frac{N(\mu - \overline{d})^2}{2\sigma^2}\right) \\
&\approx C \, \frac{\sqrt{2\pi}(\sigma/\sqrt{N})}{\Delta} \equiv C\,\Omega
\end{aligned}$$

with Ockham factor Ω

Maximum likelihood ratio

For M_2, the likelihood is maximized for $\mu = \overline{d}$ (if it's in the prior range), so the likelihood ratio is

$$R_{12} \equiv \frac{\mathcal{L}(\mu_1)}{\mathcal{L}(\overline{d})} = \exp\left(-\frac{N(\mu_1 - \overline{d})^2}{2\sigma^2}\right)$$

This is always ≤ 1 (equality only if $\mu_1 = \overline{d}$), disfavoring M_1

Bayes factor

$$B_{12} \equiv \frac{\mathcal{L}(M_1)}{\mathcal{L}(M_2)} = \frac{\Delta}{\sqrt{2\pi}\sigma/\sqrt{N}} R_{12}$$

Divides the MLR by Ockham factor (typically < 1)

$$\Omega = \frac{\sqrt{2\pi}(\sigma/\sqrt{N})}{\Delta}$$

M_2 gets penalized by the size of the μ search space, so the data may sometimes directly favor M_1

Model averaging

Problem statement

$I = (M_1 \vee M_2 \vee \ldots)$ — Specify a set of models
Models all share a set of "interesting" parameters, ϕ
Each has different set of nuisance parameters η_i (or different prior info about them)
H_i = statements about ϕ

Model averaging

Calculate posterior PDF for ϕ:

$$\begin{aligned} p(\phi|D,\mathcal{C}) &= \sum_i p(M_i|D,\mathcal{C})\, p(\phi|D,M_i) \\ &\propto \sum_i \mathcal{L}(M_i) \int d\eta_i\, p(\phi,\eta_i|D,M_i) \end{aligned}$$

The model choice is a (discrete) nuisance parameter here

Prediction

Context: Model M with parameters θ
Data: Available data D; *future data* D'
What does D tell us about D' in the context of the model?

Calculate the *posterior predictive dist'n*:

$$\begin{aligned} p(D'|D,M) &= \int d\theta\; p(\theta, D'|D,M) \\ &= \int d\theta\; p(\theta|D,M)\, p(D'|\theta,M) \\ &= \int d\theta\; (\text{posterior for } \theta) \times (\text{sampling dist'n for } D') \end{aligned}$$

Typically the last factor is easy to compute (e.g., binomial, Poisson, or normal dist'n with parameters *given*).

This is propagation of uncertainty from θ to D', by a kind of "smearing" of the sampling dist'n.

Predicting a future Bernoulli outcome

FFSSSSFSSSFS ($n = 8$ successes in $N = 12$ trials)

Bernoulli process likelihood function

$$p(\mathcal{S}|\alpha, M) = \alpha^n (1-\alpha)^{N-n}$$

Binomial likelihood function

$$p(n|\alpha, M) = \frac{N!}{n!(N-n)!}\alpha^n (1-\alpha)^{N-n}$$

Flat prior $\rightarrow$ beta dist'n posterior PDF

$$\begin{aligned} p(\alpha|n, M) &= \frac{(N+1)!}{n!(N-n)!}\alpha^n (1-\alpha)^{N-n} \\ &= \mathrm{Beta}(\alpha|a = n+1, b = N-n+1) \end{aligned}$$

Probability for next outcome

Next outcome $o = 0$ (F) or $o = 1$ (S)

$$
\begin{aligned}
p(o|n, M) &= \int d\alpha \; p(\alpha, o|n, M) \\
&= \int d\alpha \; p(\alpha|n, M)\, p(o|\alpha, M) \\
&= \int d\alpha \; \frac{(N+1)!}{n!(N-n)!} \alpha^n (1-\alpha)^{N-n} \times \alpha^o (1-\alpha)^{1-o} \\
&= \frac{(N+1)!}{n!(N-n)!} \int d\alpha \; \alpha^{n+o} (1-\alpha)^{N-n+1-o} \\
&= \frac{(N+1)!}{n!(N-n)!} \times \frac{(n+o)!(N-n-o+1)!}{(N+2)!}
\end{aligned}
$$

$$
\begin{aligned}
p(o|n,M) &= \frac{(N+1)!}{n!(N-n)!} \times \frac{(n+o)!(N-n-o+1)!}{(N+2)!} \\
&= \begin{cases} \frac{n+1}{N+2} & \text{for } o=1 \\ \frac{N-n+1}{N+2} & \text{for } o=0 \end{cases} \\
&= \frac{1}{2} \quad \text{when } N=0, n=0 \\
&\approx \begin{cases} \frac{n}{N} & \text{for } o=1 \\ \frac{N-n}{N} & \text{for } o=0 \end{cases} \quad \text{for } N,n \gg 1
\end{aligned}
$$

Laplace's rule of succession:

$P(\text{next outcome}|\text{past}) \approx \text{Frequency of outcome in the past}$

Provides justification for a simple form of inductive reasoning in IID settings

Theme: Parameter space volume

Bayesian calculations sum/integrate over parameter/hypothesis space! This is **the signature feature** of the Bayesian approach.

(Frequentist calculations average over *sample* space & typically *optimize* over parameter space.)

- Credible regions integrate over parameter space
- Uncertainty propagation integrates over parameter space
- Marginalization weights the profile likelihood by a volume factor for the nuisance parameters
- Prediction integrates over parameter space
- Model (marginal) likelihoods & Bayes factors have Ockham factors resulting from parameter space volume factors

Many/most interesting hypotheses are really *composite*. Many virtues of Bayesian methods can be attributed to accounting for the "size" of parameter spaces when considering composite hypotheses. This idea does not arise naturally in frequentist statistics (but it can be added "by hand").

Bayesian machine learning

Computer Science > Machine Learning

[Submitted on 20 Feb 2020 (v1), last revised 27 Apr 2020 (this version, v3)]

Bayesian Deep Learning and a Probabilistic Perspective of Generalization

Andrew Gordon Wilson, Pavel Izmailov

The key distinguishing property of a Bayesian approach is marginalization, rather than using a single setting of weights. Bayesian marginalization can particularly improve the accuracy and calibration of modern deep neural networks, which are typically underspecified by the data, and can represent many compelling but different solutions. We show that deep ensembles provide an effective mechanism for approximate Bayesian marginalization, and propose a related approach that further improves the predictive distribution by marginalizing within basins of attraction, without significant overhead. We also investigate the prior over functions implied by a vague distribution over neural network weights, explaining the generalization properties of such models from a probabilistic perspective. From this perspective, we explain results that have been presented as mysterious and distinct to neural network generalization, such as the ability to fit images with random labels, and show that these results can be reproduced with Gaussian processes. We also show that Bayesian model averaging alleviates double descent, resulting in monotonic performance improvements with increased flexibility. Finally, we provide a Bayesian perspective on tempering for calibrating predictive distributions.

Typical deep learning loss landscape

Machine learning tunes flexible models by *minimizing a loss function* in a high-dimensional parameter space (approximately!).

Bayesian machine learning uses

$$\text{Loss}(\theta) = -\log[\text{Posterior}(\theta)] = -\log[\text{Prior}(\theta)] - \log[\text{Likelihood}(\theta)] + C$$

and *marginalizes over the loss landscape* (very approximately!).

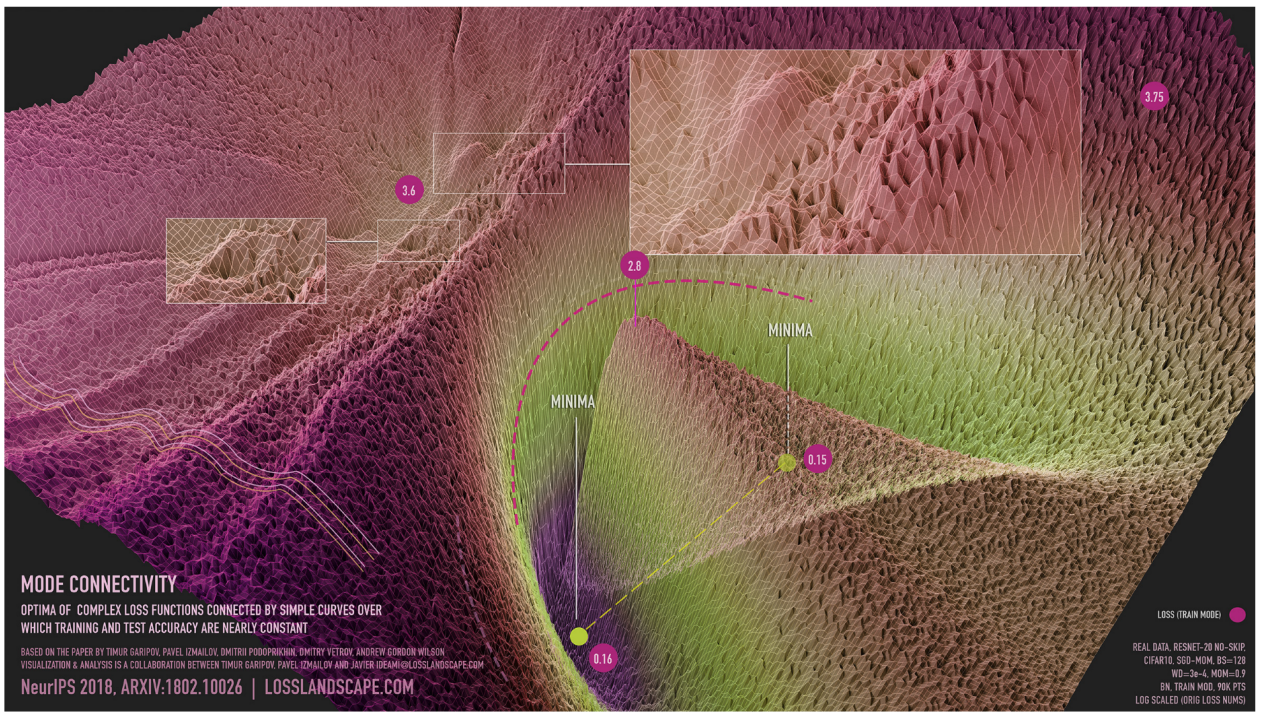


Figure 1: visualization of mode connectivity for ResNet-20 with no skip connections on CIFAR-10 dataset. The visualization is created in collaboration with Javier Ideami (https://losslandscape.com/).

Roles of the prior

Prior has two roles

- Incorporate any relevant prior information
- Convert likelihood from "intensity" to "measure"
  $\rightarrow$ account for *size of parameter space*

Physical analogy

Temperature is "intensive," heat is "extensive":

$$\text{Heat} \quad Q = \int d\vec{r}\,[\rho(\vec{r})c(\vec{r})]\,T(\vec{r})$$

$$\text{Probability} \quad P \propto \int d\theta\; p(\theta)\,\mathcal{L}(\theta)$$

Maximum likelihood focuses on the "hottest" parameters.
Bayes focuses on the parameters with the most "heat."
A high-T region may contain little heat if ρc is low or if its volume is small.
A high-$\mathcal{L}$ region may contain little probability if its prior is low or if its volume is small.