COMP90051 **Statistical Machine Learning**

Semester 2, 2016

Lecturer: Trevor Cohn

18. Bayesian classification & model selection

THE UNIVERSITY OF MELBOURNE

Recap: Bayesian inference

- Uncertainty not captured by MLE, MAP etc
- Bayesian approach preserves uncertainty
  - care about predictions NOT parameters
  - choose prior over parameters, then model posterior
  - integrate out parameters for prediction (today)
- Requires computing an integral for the 'evidence' term
  - conjugate prior makes this possible

Stages of Training

1. Decide on model formulation & prior
2. Compute *posterior* over parameters, $p(\boldsymbol{w}|\boldsymbol{X},\boldsymbol{y})$

MAP

3. Find *mode* for ***w***
4. Use to make prediction on test

approx. Bayes

3. Sample many ***w***
4. Use to make *ensemble* average prediction on test

exact Bayes

3. Use *all* ***w*** to make *expected* prediction on test

Prediction with uncertain *w*

- Could predict using sampled regression curves
  - sample S parameters, $\boldsymbol{w}^{(s)}, s \in [1, S]$
  - for each sample compute prediction $y_*^{(s)}$ at test point $\boldsymbol{x}_*$
  - compute the mean (and var.) over these predictions
  - this process is known as **Monte Carlo integration**
- For Bayesian regression there's a simpler solution
  - integration can be done analytically, giving

$$p(\hat{y}_* \mid \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{x}_*, \sigma^2) = \int p(\boldsymbol{w} \mid \boldsymbol{X}, \boldsymbol{y}, \sigma^2) p(y_* \mid \boldsymbol{x}_*, \boldsymbol{w}, \sigma^2)\, d\boldsymbol{w}$$

Prediction (cont.)

- Pleasant properties of Gaussian distribution means integration is tractable

$$p(y_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}, \sigma^2) = \int p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \sigma^2) p(y_*|\mathbf{x}_*, \mathbf{w}, \sigma^2) d\mathbf{w}$$

$$= \int \mathrm{Normal}(\mathbf{w}|\mathbf{w}_N, \mathbf{V}_N) \mathrm{Normal}(y_*|\mathbf{x}_*'\mathbf{w}, \sigma^2) d\mathbf{w}$$

$$= \mathrm{Normal}(y_*|\mathbf{x}_*'\mathbf{w}_N, \sigma_N^2(\mathbf{x}_*))$$

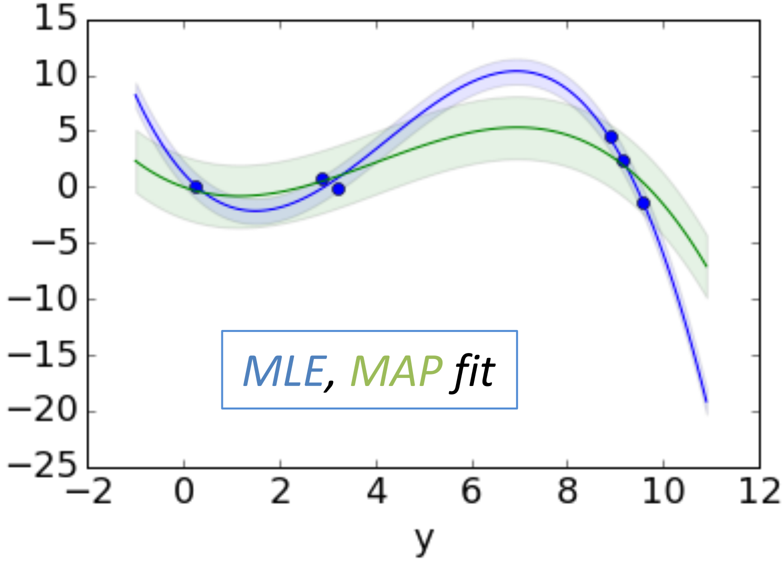
$$\sigma_N^2(\mathbf{x}_*) = \sigma^2 + \mathbf{x}_*'\mathbf{V}_N\mathbf{x}_*$$

  - additive variance based on x_* match to training data
  - cf. MLE/MAP estimate, where variance is a fixed constant

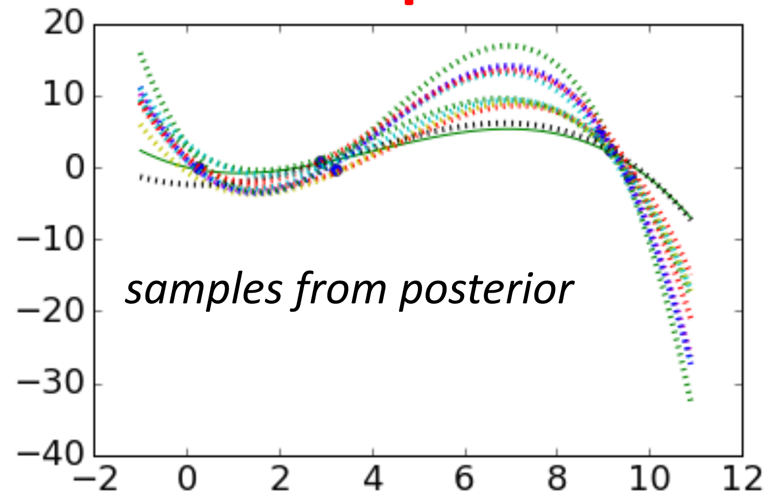
($\boldsymbol{w}_N$ and $\boldsymbol{V}_N$ defined in lecture 17)

Bayesian Prediction example

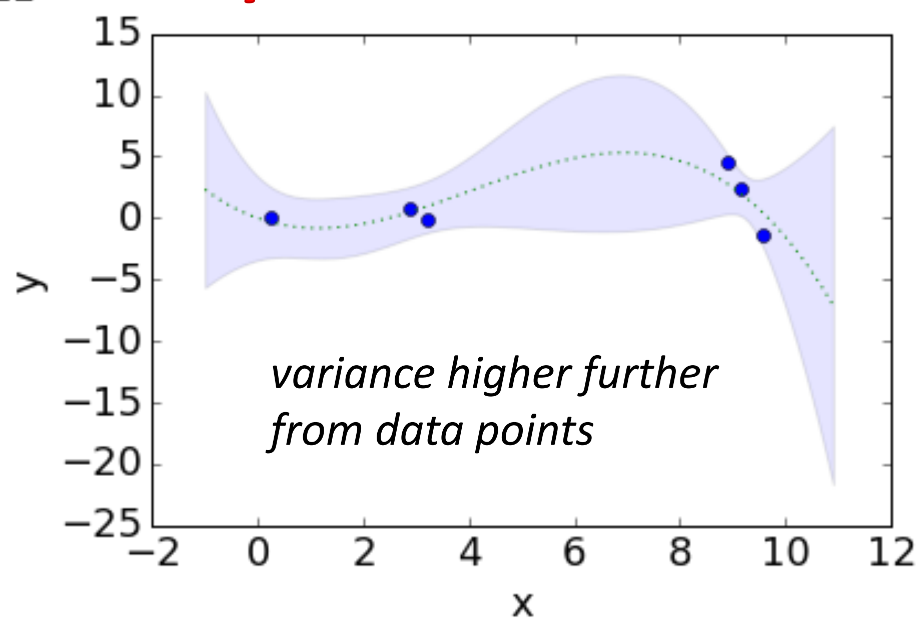
Point estimate



MLE (blue) and MAP (green) point estimates, *with fixed variance*



Bayesian inference



Data: y = x sin(x); Model = cubic

Caveats

- Assumptions
  - known data noise parameter, σ^2
  - data was drawn from the model distribution
- In real settings, σ^2 is unknown
  - has its own conjugate prior
    Normal likelihood $\times$ *InverseGamma* prior
    results in *InverseGamma* posterior
  - closed form predictive distribution, with student-T likelihood
    (see Murphy, 7.6.3)

Bayesian Classification

How can we apply Bayesian ideas to discrete settings?

Generative scenario

- First off consider models which *generate* the input
  - cf. *discriminative* models, which *condition* on the input
  - I.e., $p(y \mid \mathbf{x})$ vs $p(\mathbf{x}, y)$, Naïve Bayes vs Logistic Regression
- For simplicity, start with most basic setting
  - n coin tosses, of which k were heads
  - only have $\boldsymbol{x}$ (sequence of outcomes), but no 'classes' $\boldsymbol{y}$
- Methods apply to generative models over discrete data
  - e.g., topic models, generative classifiers (Naïve Bayes, mixture of multinomials)

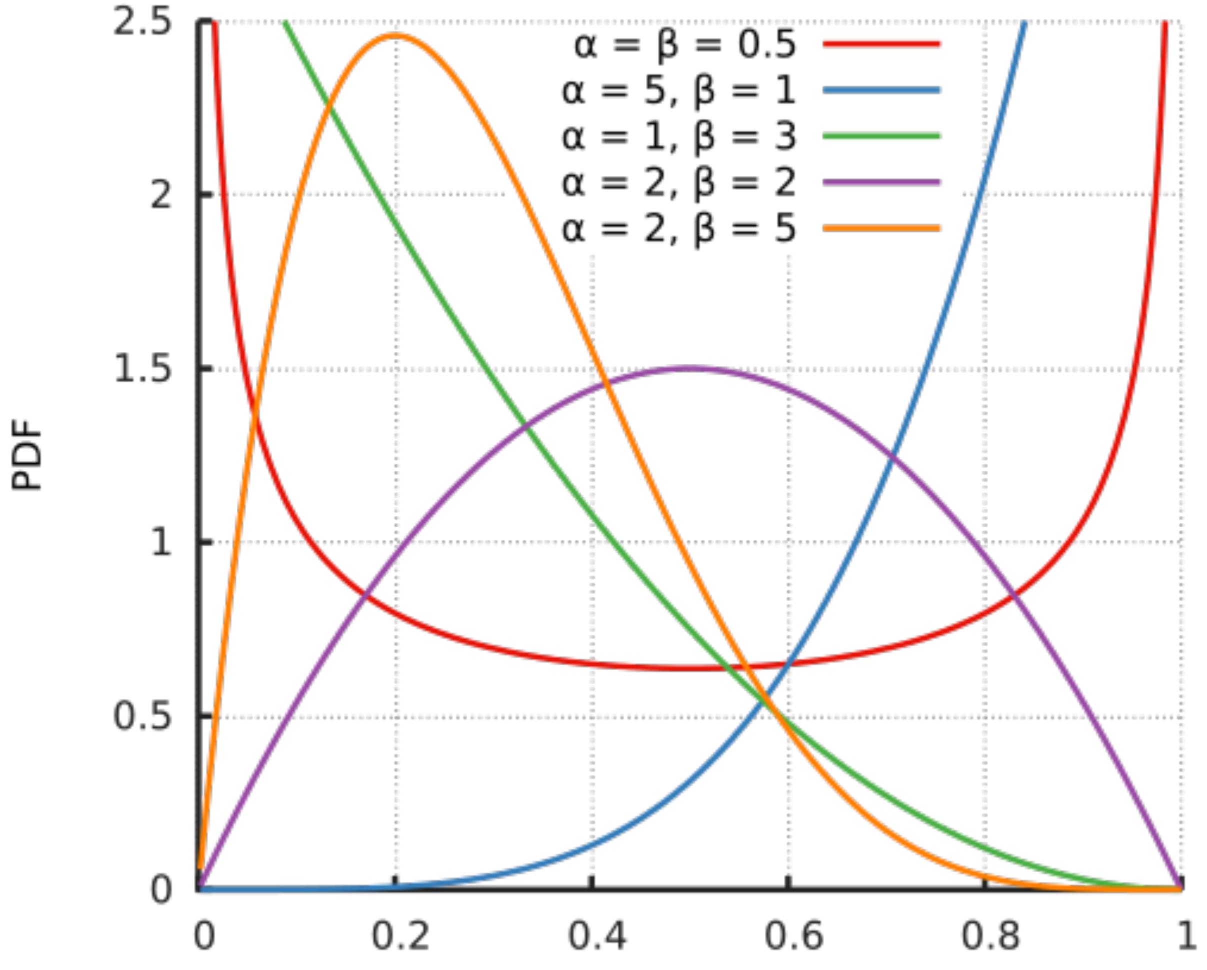
Discrete Conjugate prior: Beta-Binomial

- Conjugate priors also exist for discrete spaces
- Consider *n* coin tosses, of which *k* were heads
  - let p(head) = *q* from a single toss (*Bernoulli dist*)
  - Inference question is the coin biased, i.e., is *q ≈ 0.5*
- Several draws, use *Binomial dist*
  - and its conjugate prior, *Beta dist*

$$p(k|n,q) = \binom{n}{k} q^k (1-q)^{n-k}$$

$$p(q) = \mathrm{Beta}(q;\alpha,\beta)$$
$$= \frac{\gamma(\alpha+\beta)}{\gamma(\alpha)\gamma(\beta)} q^{\alpha-1}(1-q)^{\beta-1}$$

Beta distribution



Sourced from https://en.wikipedia.org/wiki/Beta_distribution

Beta-Binomial conjugacy

$$p(k|n,q) = \binom{n}{k} q^k (1-q)^{n-k}$$

$$p(q) = \mathrm{Beta}(q;\alpha,\beta)$$

$$= \frac{\gamma(\alpha+\beta)}{\gamma(\alpha)\gamma(\beta)} q^{\alpha-1}(1-q)^{\beta-1}$$

Sweet! We known the normaliser for Beta

Bayesian posterior

$$p(q|k,n) \propto p(k|n,q)p(q)$$

$$\propto q^k(1-q)^{n-k} q^{\alpha-1}(1-q)^{\beta-1}$$

$$= q^{k+\alpha-1}(1-q)^{n-k+\beta-1}$$

$$\propto \mathrm{Beta}(q; k+\alpha, n-k+\beta)$$

trick: ignore constant factors (normaliser)

Laplace's Sunrise Problem

Every morning you observe the sun rising. Based solely on this fact, what's the probability that the sun will rise tomorrow?

- Use beta-binomial, where *q* is the Pr(sun rises in morning)
  - posterior $$p(q|k,n) = \mathrm{Beta}(q; k+\alpha, n-k+\beta)$$
  - n = k = age in days
  - let alpha = beta = 1 (*uniform* prior)
- Under these assumptions

$$p(q|k) = \mathrm{Beta}(q; k+1, 1)$$

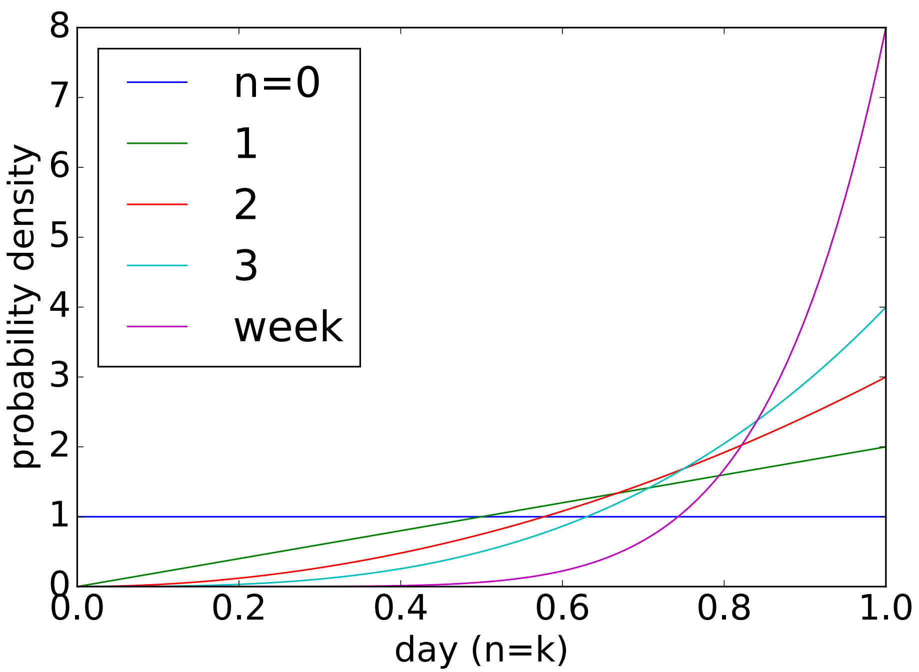
$$E_{p(q|k)}[q] = \frac{k+1}{k+2}$$

'smoothed' count of days where sun rose / did not

Sunrise Problem (cont.)

Consider a human life-span

Day (n, k)	k+α	n-k+β	E[q]
0	1	1	0.5
1	2	1	0.667
2	3	1	0.75
…			
365	366	1	0.997
2920 (80 years)	2921	1	0.99997



Effect of prior diminishing with data, *but never disappears completely*.

Suite of useful conjugate priors

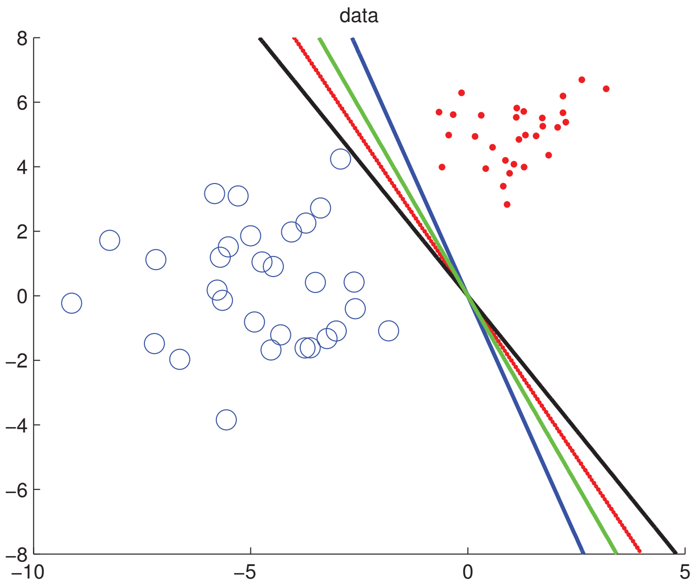
	likelihood	conjugate prior
regression	Normal	Normal (for mean)
	Normal	Inverse Gamma (for variance) or Inverse Wishart (covariance)
classification	Binomial	Beta
	Multinomial	Dirichlet
counts	Poisson	Gamma

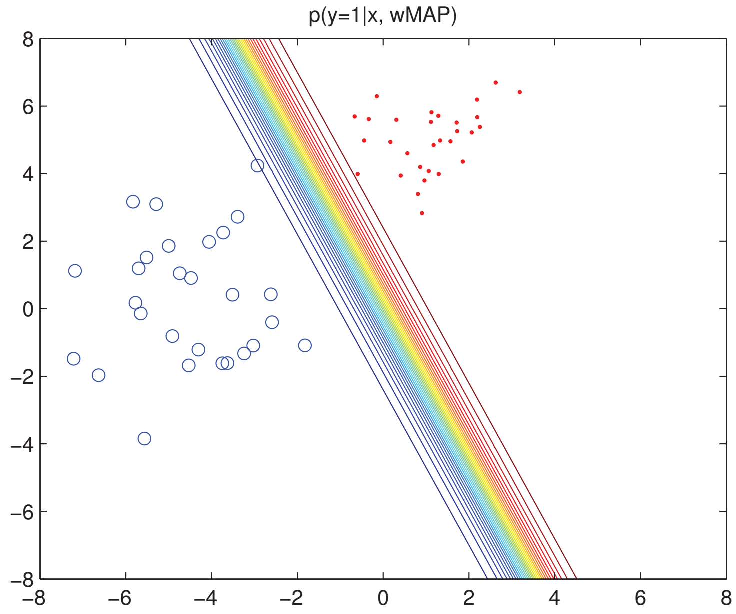
Bayesian Logistic Regression

Discriminative classifier, which *conditions* on inputs. How can we do Bayesian inference in this setting?

Now for Logistic Regression…

- Similar problems with parameter uncertainty compared to regression
  - * although predictive uncertainty in-built to model outputs





Murphy Fig 8.5 & 8.6 p257-8

Now for Logistic Regression…

- Can we use conjugate prior? E.g.,
  - Beta-Binomial for *generative* binary models
  - Dirichlet-Multinomial for multiclass (similar formulation)
- Model is *discriminative*, with parameters defined using logistic sigmoid*

$$p(y|q,\mathbf{x}) = q^y(1-q)^{1-y}$$
$$q = \sigma(\mathbf{x}'\mathbf{w})$$

  - need prior over ***w***, not *q*
  - no known conjugate prior (!), thus use a Gaussian prior

** Or softmax for multiclass; same problems arise and similar solution*

Non-conjugacy

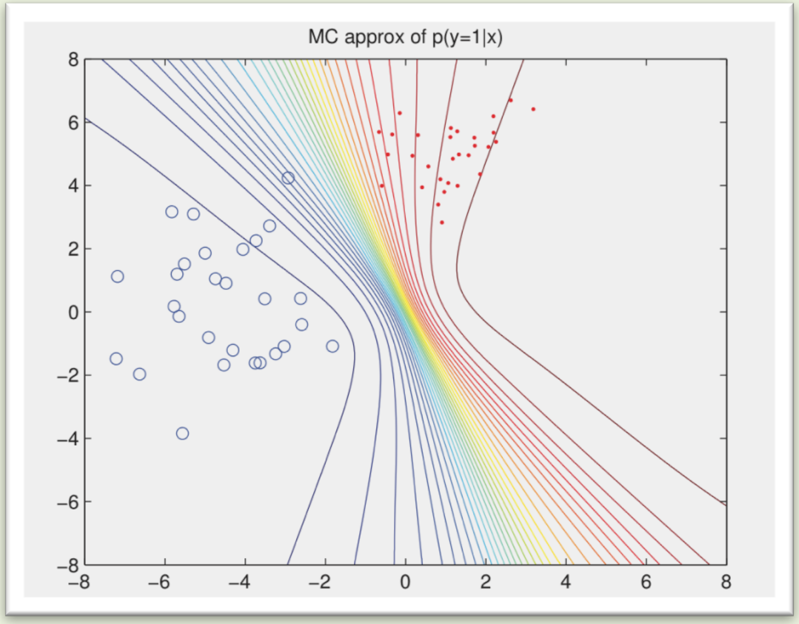
- No known solution for the normalising constant

$$p(\mathbf{w}|\mathbf{X},\mathbf{y}) \propto p(\mathbf{w})p(\mathbf{y}|\mathbf{X},\mathbf{w})$$

$$= \text{Normal}(\mathbf{0}, \sigma^2\mathbf{I}) \prod_{i=1}^{n} \sigma(\mathbf{x}_i'\mathbf{w})^{y_i}(1-\sigma(\mathbf{x}_i'\mathbf{w}))^{1-y_i}$$

- Resolve by *approximation*

***Laplace approx.*:**
- assume posterior ≃ Normal about mode
- can compute normalisation constant, draw samples etc.



Murphy Fig 8.6 p258

Bayesian Model Selection

Using the *evidence* to select the best *class* of model.

Model Selection

- Choosing the best model
  * linear, polynomial order, RBF basis/kernel
  * setting model hyperparameters
  * optimiser settings
  * type of model (e.g., decision tree vs svm)

Complex models:
- better ability fit the training data
- may fit it too well

Simple models:
- more constrained, poorer fit to training data
- might be insufficient

Model Selection (frequentist)

- *Holdout* some data for validation (fixed set, leave-one-out, 10-fold cross valid., etc)
  - treat held-out error as estimate of generalisation error
  - model with lowest error is chosen
  - might retrain chosen model on full dataset
- However, this is
  - data inefficient: must hold aside evaluation data
  - computationally inefficient: repeatedly rounds of training and evaluating
  - ineffective: when selecting many parameters at once (can overfit the heldout set)

Bayesian Model Selection

- Model selection using Bayes rule, to select between competing model classes

$$p(\mathcal{M}_i|\mathcal{D}) = \frac{p(\mathcal{D}|\mathcal{M}_i)p(\mathcal{M}_i)}{p(\mathcal{D})}$$

  * with M_i as model i and D the dataset e.g., $\boldsymbol{X}$ or $\boldsymbol{y}|\boldsymbol{X}$, so for regression $p(D) = p(\boldsymbol{y}|\boldsymbol{X})$
  * let $p(M_i)$ be uniform; i.e., term dropped

- Decision between two model classes boils down to test (known as Bayes factor)

$$\frac{p(\mathcal{M}_1|\mathcal{D})}{p(\mathcal{M}_2|\mathcal{D})} = \frac{p(\mathcal{D}|\mathcal{M}_1)}{p(\mathcal{D}|\mathcal{M}_2)} > 1$$

The Evidence: *p(D|M) = p(**y**|**X**,M)*

- Imagine we're considering whether to use a linear basis or cubic basis for supervised regression
  - what is $p(\mathbf{y}|\mathbf{X}, M=\text{linear})$ or $p(\mathbf{y}|\mathbf{X}, M=\text{cubic})$?
  - what happened to the parameters **w**?
- These are integrated out, i.e.,

$$p(\mathbf{y}|\mathbf{X}, M = \text{linear}) = \int p(\mathbf{y}|\mathbf{X}, \mathbf{w}, M = \text{linear})p(\mathbf{w})d\mathbf{w}$$

  - seen before: the denominator from posterior, aka '*marginal likelihood*'

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}, M) = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w}, M)p(\mathbf{w})}{p(\mathbf{y}|\mathbf{X}, M)}$$

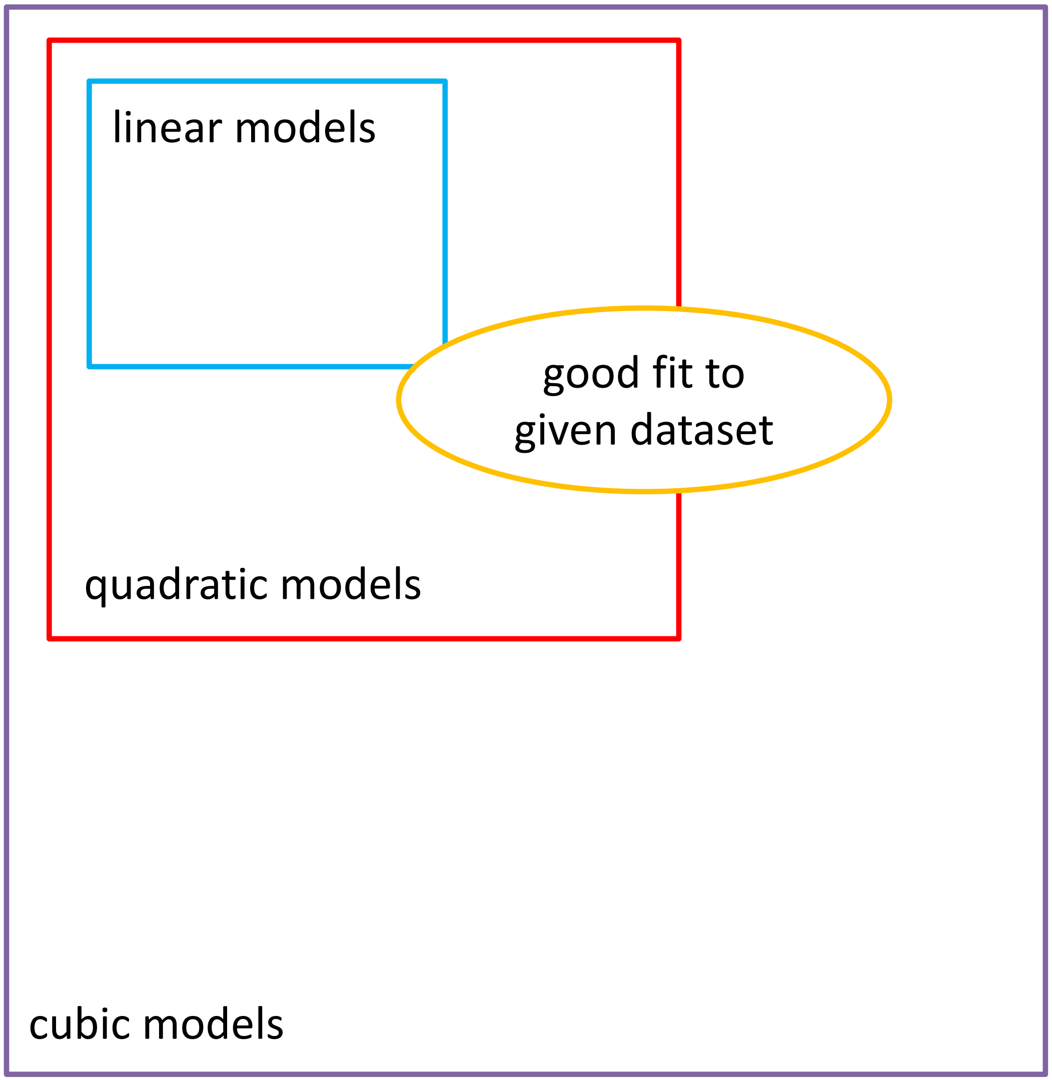
The Evidence: Bayesian Occam's Razor

- How well does the model fit the data, *under any (all) parameter settings?*

- Flexible (complex) models
  - able to fit many different datasets, by selecting specific parameters
  - most other parameter settings will lead to a poor fit

- Simpler models
  - fit few datasets well
  - less sensitive to parameter values
  - many parameter settings will give similar fit

Evidence Cartoon (under uniform prior)



- Space of models: linear < quadratic < cubic
- *Assuming quadratic data*, this is best fit by >= quadratic model
- As complexity class grows, space of models grows too
  - fraction of params offering ‘good’ fit to data will shrink
- Ideally, would select quadratic model as fraction is greatest

Summary

- Conjugate prior relationships
  - Normal-Normal, Beta-Binomial
- Bayesian inference
  - parameters are ‘nuisance’ variables
  - integrated out during inference
- Bayesian classification
  - non-conjugacy necessitates approximation
- Bayesian model selection