# Lecture 19. Bayesian classification

COMP90051 Statistical Machine Learning

Lecturer: Zahra Dasht Bozorgi

THE UNIVERSITY OF MELBOURNE

Copyright: University of Melbourne

# This lecture

- Bayesian ideas in discrete settings
  - Beta-Binomial conjugacy
  - Uniqueness up to proportionality
  - Sunrise example
  - Common conjugate pairs
- Bayesian logistic regression
  - Non-conjugacy
  - Pointer: Laplace approximation

# How to apply Bayesian view to discrete data?

- First off consider models which *generate* the input
  - cf. *discriminative* models, which *condition* on the input
  - I.e., $p(y \mid \mathbf{x})$ vs $p(\mathbf{x}, y)$, Logistic Regression vs Naïve Bayes
- 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

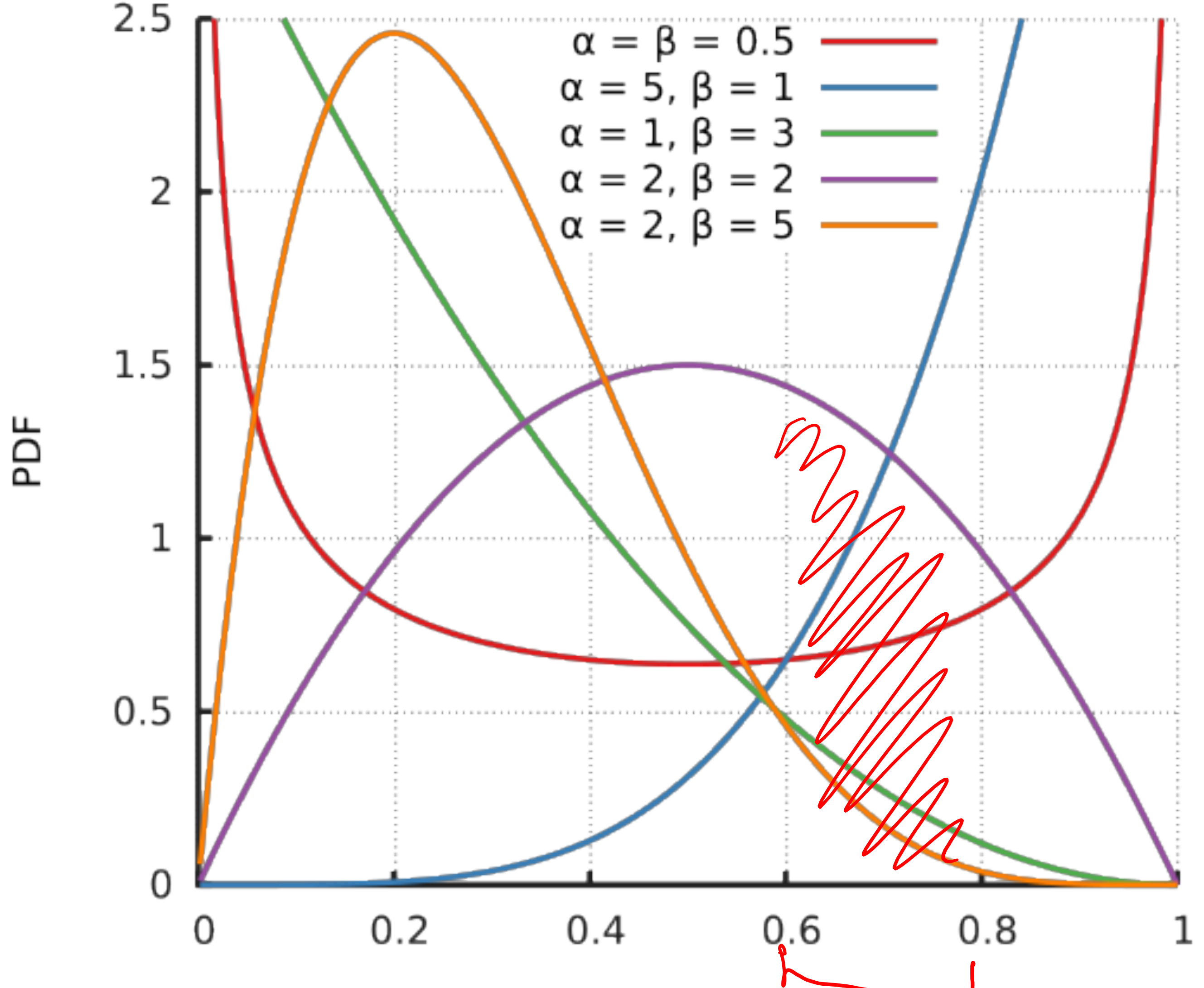
- Conjugate priors also exist for discrete spaces
- Consider *n* coin tosses, of which *k* were heads
  - let p(head) = *q* from a single toss (*Bernoulli dist*)
  - Inference question is the coin biased, i.e., is *q ≈ 0.5*
- Several draws, use *Binomial dist*
  - and its conjugate prior, *Beta dist*

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

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

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

# Beta distribution




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

# Beta-Binomial conjugacy

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

likehood

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

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

normalising term.

Sweet! We know 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)

# Uniqueness up to normalisation

- A trick we've used many times:

  *When an unnormalized distribution is proportional to a recognised distribution, we say it must be that distribution*

- If $f(\theta) \propto g(\theta)$ for $g$ a distribution, $\dfrac{f(\theta)}{\int_\Theta f(\theta)d\theta} = g(\theta)$.

- <u>Proof</u>: $f(\theta) \propto g(\theta)$ means that $\exists C$

$$f(\theta) = C \cdot g(\theta)$$

$$\int_\Theta f(\theta)d\theta = C \int_\Theta g(\theta)d\theta = C$$

and the result follows from LHS1/LHS2 = RHS1/RHS2

# 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?*

每日升起是不独立的.

- Use Beta-Binomial, where *q* is the Pr(sun rises in morning)
  - posterior $$p(q|\underbrace{k,n}_{\#\text{of obs}}) = \text{Beta}(q; k+\alpha, n-k+\beta)$$
  - n = k = observer's age in days
  - let $\alpha = \beta = 1$ (*uniform* prior)

    Beta(1,1) = uniform

likelihood = ?

prior = ?

$\theta^N$ = likelihood.

- Under these assumptions

$$p(q|k) = \text{Beta}(q; \underbrace{k+1}_{\text{加平滑}}, 1)$$

连续k天 → k+1

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

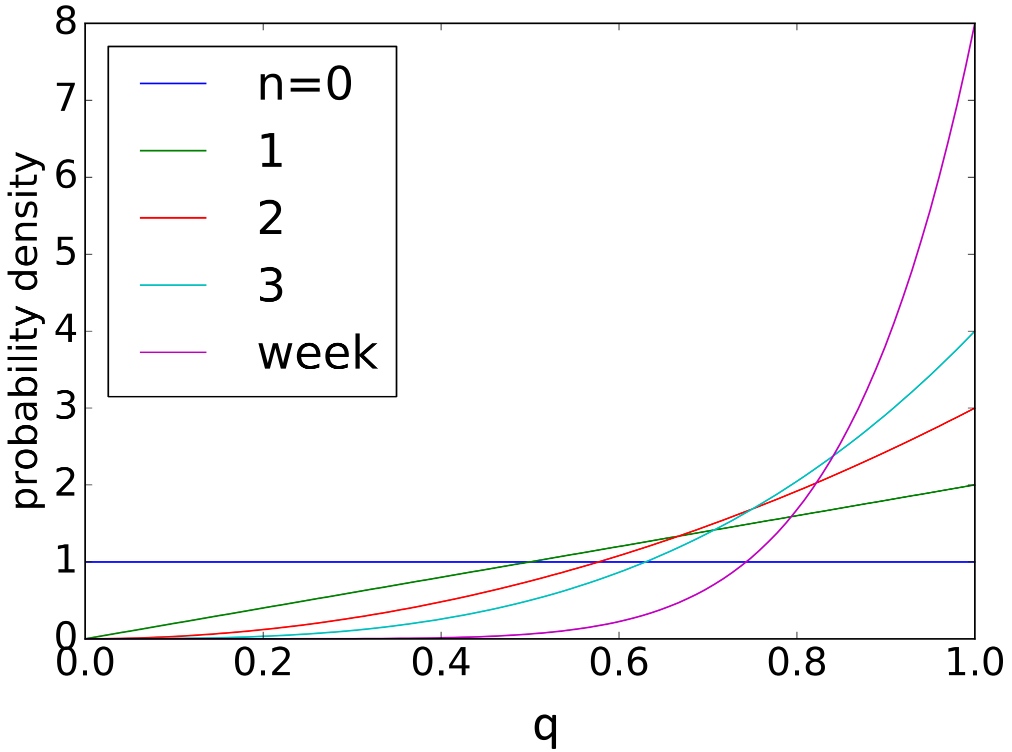
不平滑 Beta(k,1) ⇒ $\frac{k}{k+1}$ (小样本估计时极难给判断

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

# Sunrise Problem (cont.)

## Consider human-meaningful period

| 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 (8 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)<br>or Inverse Wishart (covariance) |
| classification | Binomial | Beta |
| | Multinomial | Dirichlet |
| counts | Poisson | Gamma |

# Mini Summary

- Bayesian ideas in discrete settings
  - Beta-Binomial conjugacy
  - Uniqueness in proportionality
  - Sunrise example
  - Conjugate pairs

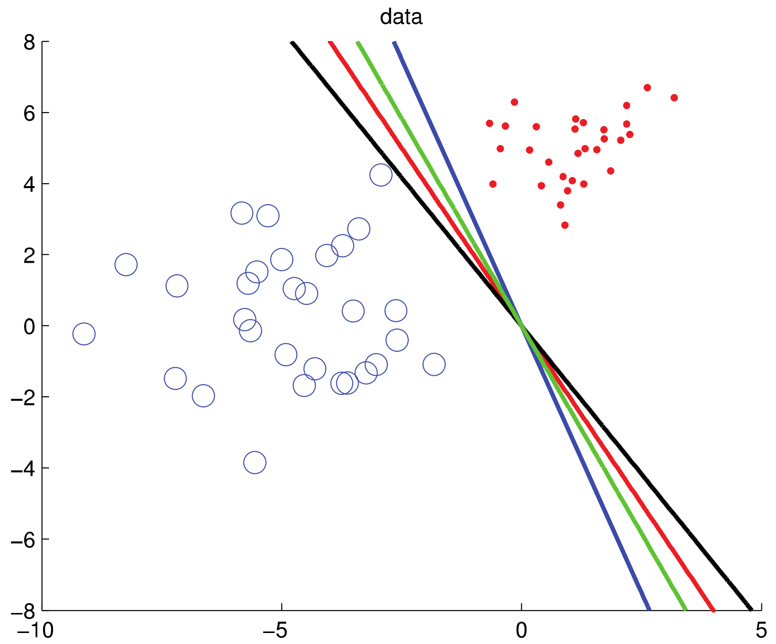
Next time: Bayesian logistic regression

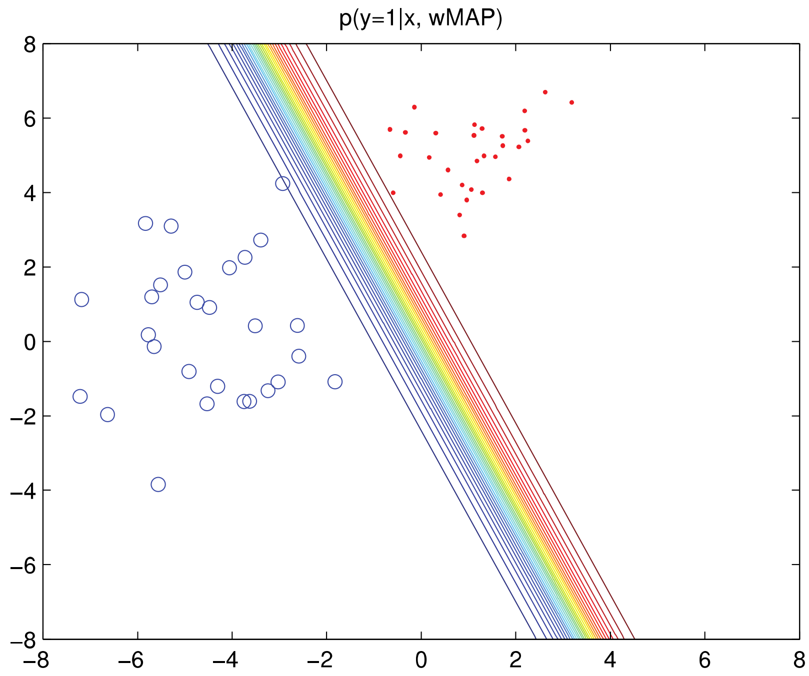
# Bayesian Logistic Regression

*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

# No conjugacy

- 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
- Approach to inference: Monte Carlo sampling

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

# Approximation

- No known solution for the normalising constant

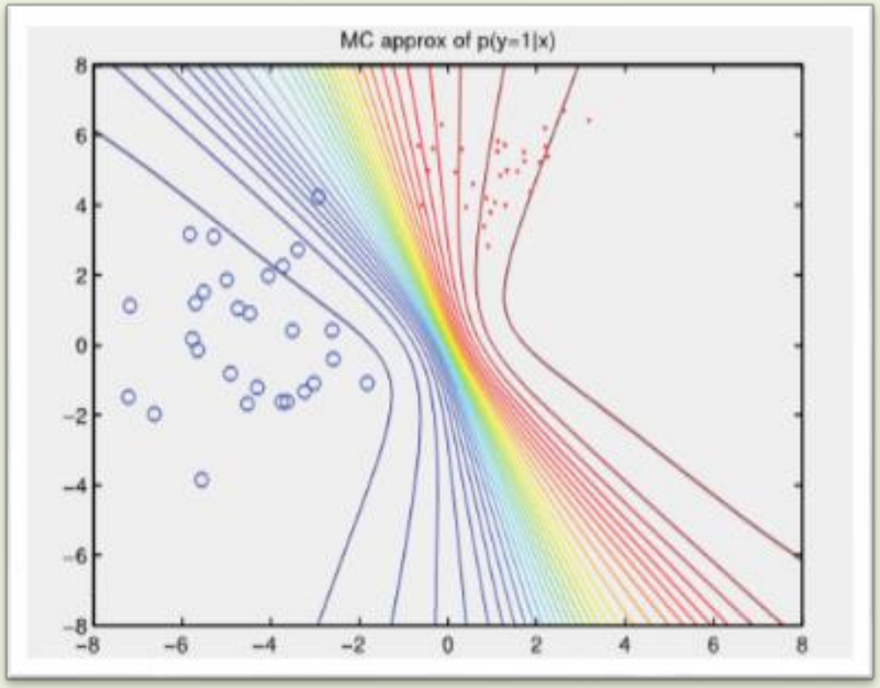
$$p(\mathbf{w}|\mathbf{X},\mathbf{y}) \propto p(\mathbf{w})p(\mathbf{y}|\mathbf{X},\mathbf{w})$$

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

- Resolve by *approximation*

***Laplace approx.*:**

- assume posterior ≃ Normal about mode
- can compute normalisation constant, draw samples etc.
- Tractable MAP provides parameters for this (Normal) approximate posterior




Murphy Fig 8.6 p258

# Mini Summary

- Bayesian ideas in discrete settings
  - Beta-Binomial conjugacy
  - Conjugate pairs; Uniqueness in proportionality
- Bayesian classification (logistic regression)
  - Non-conjugacy necessitates approximation

Next time: probabilistic graphical models