Vanilla Bayesian inferential & predictive paradigm

$P = (Q, C)$ (natural language) $\mapsto (\underset{\sim}{\theta}, \mathcal{D}, B)$ (math)

AM disc. sec. (1)

$\rightarrow M_{IP} = \{ p(\underset{\sim}{\theta} | B),\ p(\mathcal{D} | \underset{\sim}{\theta} B) \rightarrow \ell(\underset{\sim}{\theta} | \mathcal{D} B) \}$

$\underset{\sim}{\theta} \in \Theta$ (parameter space)

input single choice for prior, single choice for sampling distribution/likelihood, use Bayes's Theorem to get posterior $p(\underset{\sim}{\theta} | \mathcal{D} B) =$ optimal inference given M_{IP}

$c\, p(\underset{\sim}{\theta} | B)\, \ell(\underset{\sim}{\theta} | \mathcal{D} B)$

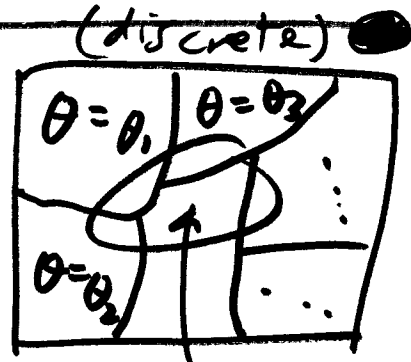
Bayesian prediction: new data $\mathcal{D}^*$, not yet observed: let's predict $\mathcal{D}^*$ based on context, prior info, likelihood info from $\mathcal{D}$ (already-observed data)

posterior predictive distribution for $\tilde{D}^*$ given D, $\mathcal{C}$, prior info

$$p(\tilde{D}^* \mid D\,\mathcal{B}) = ? \quad \text{(hard)}$$

(prior) (IM$_{IP}$)

as with de Finetti's theorem, the problem becomes easier if we know $\tilde{\theta}$, so let's bring $\tilde{\theta}$ into the calculation by partitioning over it:



$$p(\tilde{D}^* \mid D\,\mathcal{B}) = \int_\Theta p(\tilde{D}^* \tilde{\theta} \mid D\,\mathcal{B})\, d\theta$$

$$= \int_\Theta \underbrace{p(\tilde{D}^* \mid \tilde{\theta}\, \cancel{D}\, \mathcal{B})}\; p(\tilde{\theta} \mid D\,\mathcal{B})\, d\theta$$

this is just the Law of Total Probability for continuous θ

Now something excellent happens:

ex. $(\tilde{Z}_i \mid \tilde{\theta}\,\mathcal{B}) \overset{\text{IID}}{\sim} \text{Bernoulli}(\theta)$ $(i = 1, \ldots, n)$ (model $\mathcal{B}$)

If I know θ, the $\tilde{Z}_i$ are conditionally independent given θ

Bayesian interpretation / definition of ③ independence: X_1, X_2 independent if information about X_1 doesn't change your (predictive) distribution for X_2 & vice versa

in model ⊛, if we know θ, there's no information in $(\Sigma_1, \ldots, \Sigma_n)$ for predicting Σ_{n+1} but if we don't know θ, $(\Sigma_1, \ldots, \Sigma_n)$ has lots of information in it for predicting Σ_{n+1} (ex. in first $n=10$ data values, $5=8$ were 1s; you would predict $\Sigma_{n+1} =$)

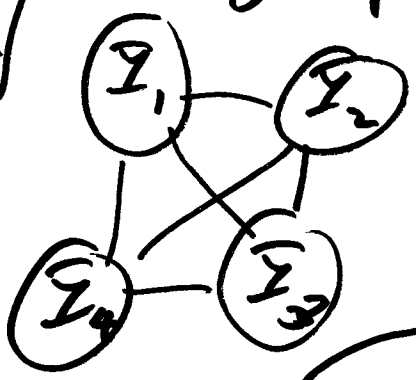
connection with Mr. deFinetti / before we saw

the data $(\underset{\sim}{I} = \underset{\sim}{i})$: $(I_1 = i_1, I_2 = i_2, \ldots, I_n = i_n)$, our uncertainty is exchangeable, but exchangeability $\neq$ (I ~~I~~ I): we just agree that there's information in any of the I_j that's helpful in predicting any of the other I_i, so exchangeable I_j are not independent; but in this example they become independent if we know θ:

exchangeability = conditionally IID given θ

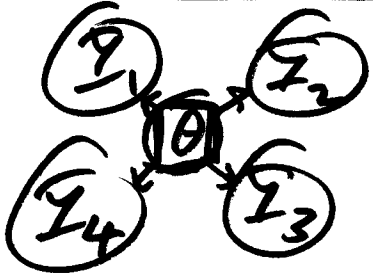
ex.

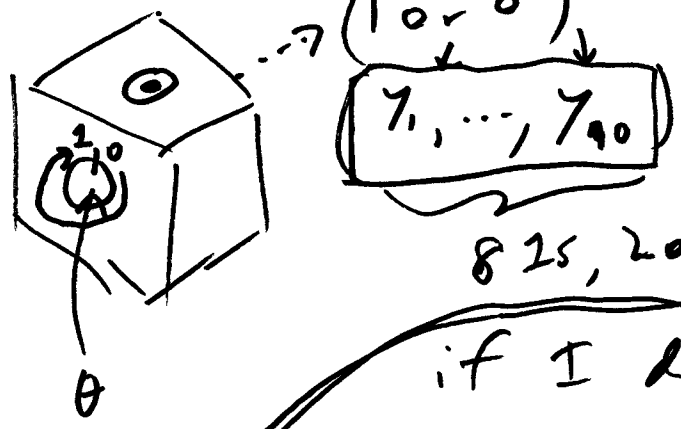
graph



$\frac{n(n-1)}{2}$ connections

$n = 4$ nodes, $\binom{n}{2} = 6 = \Omega(n^2)$





$(Y_i \mid \theta, \beta) \overset{IID}{\sim} \text{Bernoulli}(\theta)$

$(i = 1, \ldots, n)$

$\{15, 20\}$ $(n = 10)$

if I don't know θ, there

(i) useful information in

$(Y_1, \ldots, Y_{10})$ for predicting Y_{11}:

with $(8\,15, 2\,05)$ in $(Y_1, \ldots, Y_{10})$

you'd predict $Y_{11} = 1$

if θ is unknown

the $(Y_1, \ldots, Y_{11})$ are dependent

$\neq$ IID ✓

but exchangeable

now instead,

(ii) θ known $= 0.87$

$(9\,15, 1\,00)$: if I know θ,

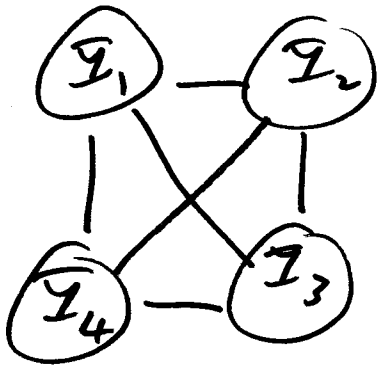
there's no info in that helps

me predict Y_{11}; if θ is known,

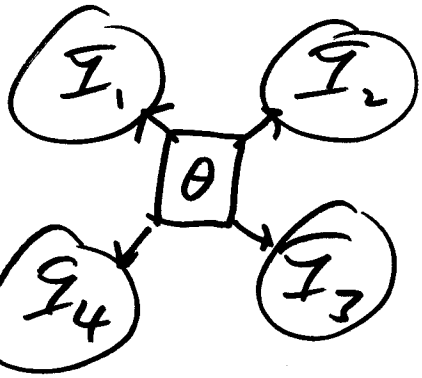
$(Y_1, \ldots, Y_{11})$ are now (conditionally independent given θ

graph theory

(unknown θ)



known θ



(nodes)

○ (unknown)

□ (known)

$n = 4 = O(n)$ nodes

$(n+1) = 5$

~~$n = 5$~~ $= \underline{\underline{O}}(n)$ nodes

$\binom{n}{2} = \binom{4}{2}$

$= \frac{4 \cdot 3}{2 \cdot 1}$

$= 6$ edges

$\binom{n}{1} = n = \underline{\underline{O}}(n)$ edges

○—○ dependent

$\theta \rightarrow I_1$ "θ causes I_1"

$= \frac{n(n-1)}{2}$

$= \underline{\underline{O}}(n^2)$

(complexity of the graph)

conditional independence given θ dramatically reduces probabilistic model complexity

exchangeability

optimal prediction

$p(D^* | \underline{D}\, B)$ | $P = (Q, \mathcal{A}) \to (\underline{\theta}, D, B)$

posterior predictive distribution

given D, prior info, M_{EP}, B

$$p(D^* | D B) = \int_{\Theta} p(D^* | \underline{\theta}\, B) \cdot p(\underline{\theta} | D B)\, d\underline{\theta}$$

l.s case study

Lecture notes part 6 pp. 7 → 12

$\underline{Z} = (Y_1, \ldots, Y_n)$

$(Y_i | \lambda P B) \overset{IID}{\sim} \text{Poisson}(\lambda)$ ($\lambda > 0$)

($i = 1, \ldots, n$)

discrete

PMF for Y_i

$$p(Y_i = y_i | \lambda P B) = \frac{\lambda^{y_i} e^{-\lambda}}{y_i!} I(y_i = 0 \text{ or } 1 \text{ or } \ldots)$$

PDF for λ

$$q(\lambda | \underline{z} P B) = c\, \lambda^{s} e^{-n\lambda}, \quad s = \sum_{i=1}^{n} y_i$$

(conjugate prior) $p(\lambda | \alpha, \beta) = \Gamma(\alpha, \beta)$ ← gamma $(\alpha > 0, \beta > 0)$

$$p(\lambda | \alpha, \beta) = c\, \lambda^{\alpha-1} e^{-\beta\lambda} I(\lambda > 0)$$

---

prior sample size = β

prior mean (estimate of λ) $= \frac{\alpha}{\beta} = \mu$

$\alpha = \beta\mu$

LI prior: $\beta = \epsilon > 0$ ← small (e.g.) positive (0.001)

class: $\mu = 5$

4toh: $\mu = 16$

2 main uses for predictive dist.

① genuine interest in predicting new data

② model-checking (LOO) leave one out = jackknife