CS340 Machine learning
Midterm review

Topics

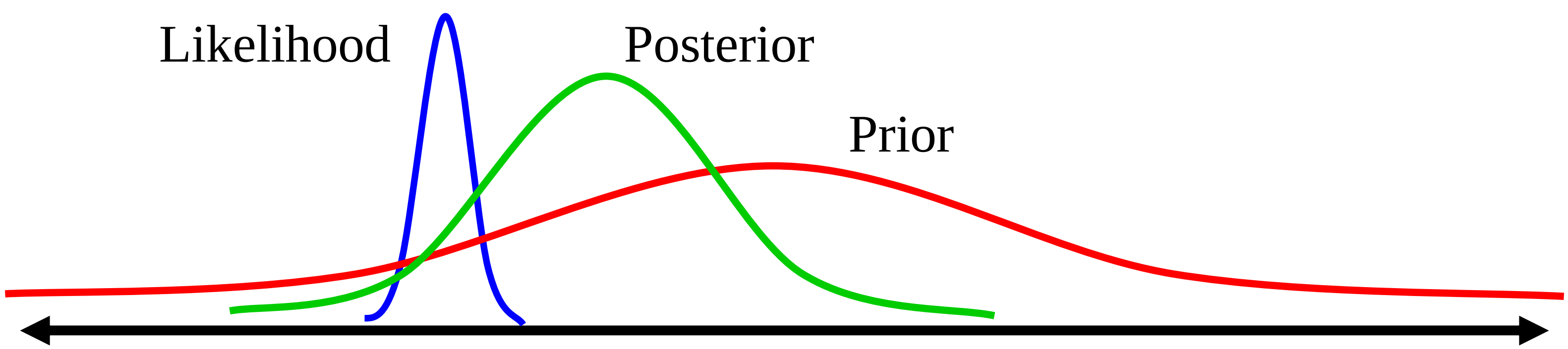
→ • Bayesian statistics
- Information theory
- Decision theory

kNN not on exam
Sampling distributions (confidence intervals etc) not on exam

Bayesian belief updating

Posterior probability — Likelihood — Prior probability

$$p(h \mid d) = \frac{p(d \mid h)\, p(h)}{\sum_{h' \in H} p(d \mid h')\, p(h')}$$



Bayesian inference = Inverse probability theory

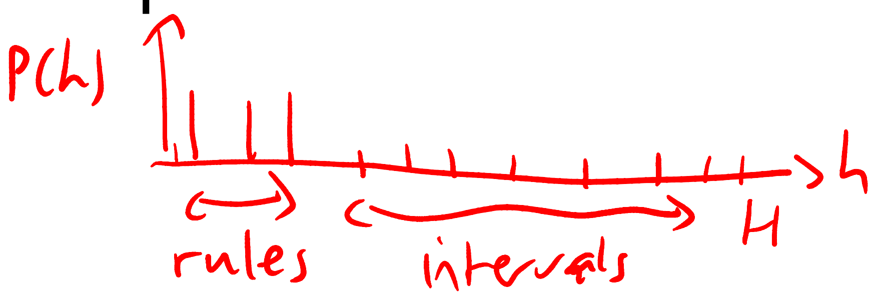
Number game

- Data: $x_i \in \{1,\ldots,100\}$.
- Hypothesis space: $h \in \{1, \ldots, H\}$ H ~ 3000

  (rules + intervals)
- Likelihood: strong sampling model

$$p(D|h) = \left[\frac{1}{|h|}\right]^n I(x_1, \ldots, x_n \in h)$$

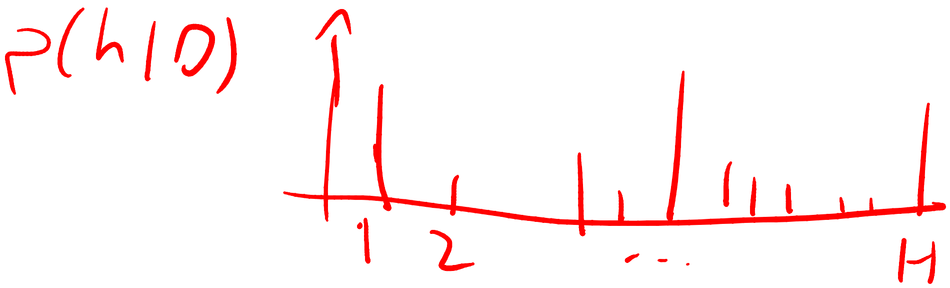
- Prior: piecewise uniform histogram



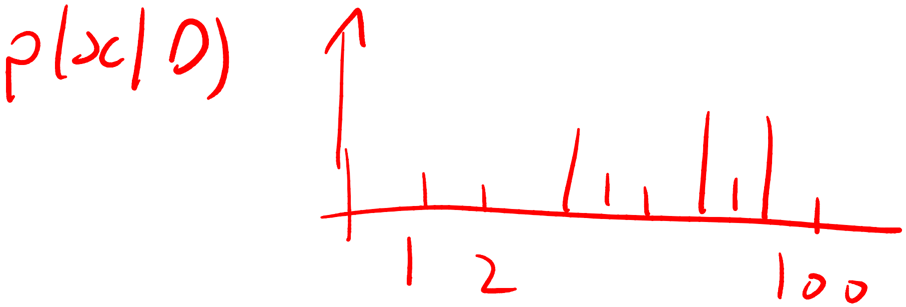
$$p(h) = \frac{\lambda I(h \in \text{rules}) + (1-\lambda) I(h \in \text{intervals})}{H}$$

Number game

- Posterior: histogram



- Posterior predictive: histogram



Coin tossing

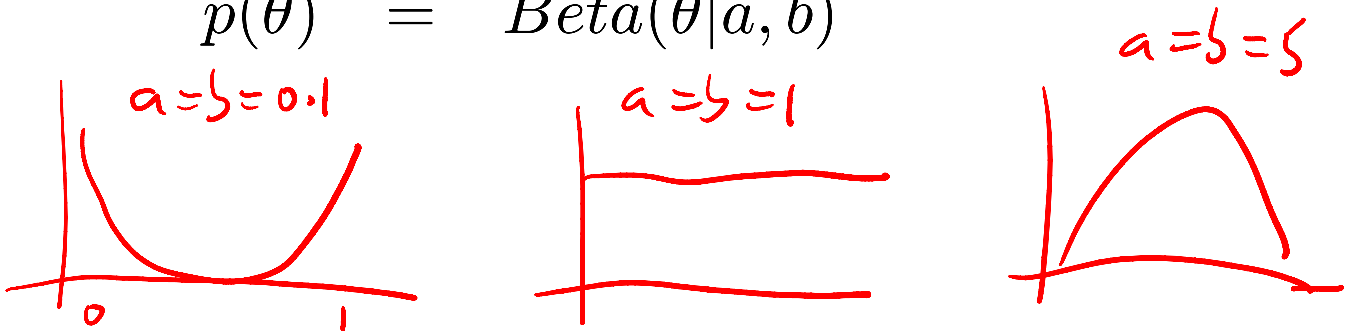
- Data: $x_i \in \{0,1\}$
- Hypothesis space: θ in [0,1]
- Likelihood: Bernoulli

$$
\begin{aligned}
p(D|\theta) &= \prod_{i=1}^{n} \theta^{I(x_i=1)}(1-\theta)^{I(x_i=0)} = \theta^{N_1}(1-\theta)^{N_0} \\
N_j &= \sum_{i=1}^{n} I(x_i = j)
\end{aligned}
$$

$$E[\theta] = \frac{a}{a+b}$$

- Prior: Beta

$$p(\theta) = Beta(\theta|a,b)$$



Coin tossing

- Posterior: beta

$$p(\theta|D) \quad = \quad Beta(\theta|a + N_1, b + N_0)$$

- Posterior predictive: two-element histogram

$$\begin{aligned} p(X = 1|D) \quad &= \quad \int p(X = 1|\theta)p(\theta|D)d\theta \\ &= \quad \int \theta Beta(\theta|a', b')d\theta = \frac{a'}{a' + b'} \end{aligned}$$

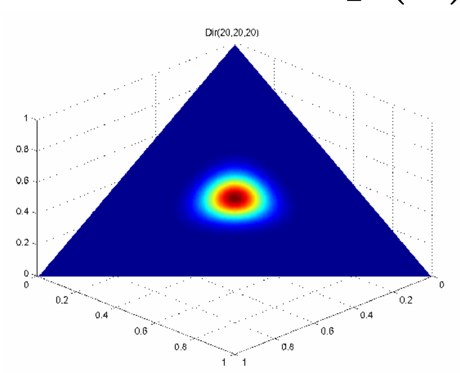
Dice rolling

- Data: $x_i \in \{1,\ldots,K\}$
- Hypothesis space: $(\theta_1, \ldots, \theta_K) \in [0,1]^K$ st $\sum_k \theta_k=1$ (probability simplex)
- Likelihood: Multinomial

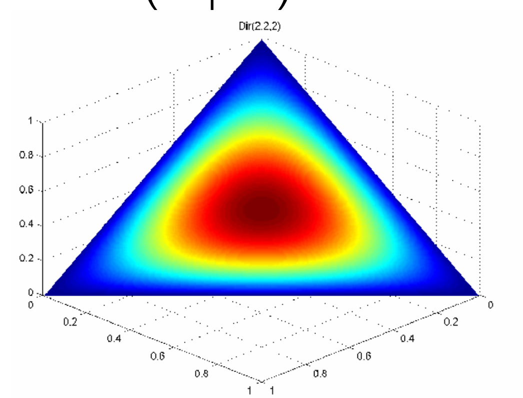
$$p(D|\boldsymbol{\theta}) = \prod_k \theta_k^{N_k}$$

- Prior: Dirichlet

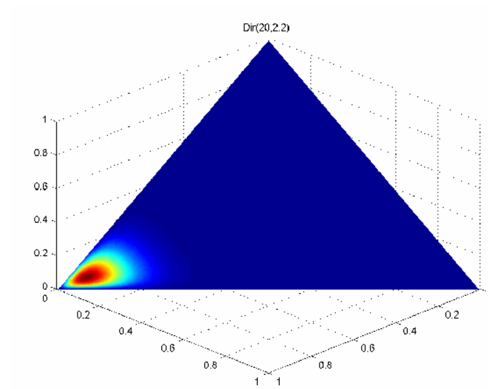
$$p(\boldsymbol{\theta}) = Dir(\boldsymbol{\theta}|\boldsymbol{\alpha})$$



(20,20,20)



(2,2,2)



(20,2,2)

Dice rolling

- Posterior: Dirichlet

$$p(\boldsymbol{\theta}|D) \quad = \quad Dir(\boldsymbol{\theta}|\boldsymbol{\alpha} + \mathbf{N})$$

- Posterior predictive: K-element histogram

$$p(X = k|D) \quad = \quad E[\theta_k|D] = \frac{\alpha_k + N_k}{\sum_{k'} \alpha_{k'} + N_{k'}}$$

Real values

- Data: $x_i \in R$
- Hypothesis space: $\mu \in R$ (λ known)
- Likelihood: Gaussian

$$p(D|\mu) = \frac{1}{(2\pi)^{n/2}}\lambda^{n/2}\exp\left(-\frac{\lambda}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right)$$

$$= \frac{1}{(2\pi)^{n/2}}\lambda^{n/2}\exp\left(-\frac{\lambda}{2}\left[n(\mu-\overline{x})^2+\sum_{i=1}^{n}(x_i-\overline{x})^2\right]\right)$$

- Prior: Gaussian

$$p(\mu) = \mathcal{N}(\mu|\mu_0, \lambda_0^{-1})$$

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n}x_i$$

Real values

- Posterior: Gaussian

$$
\begin{aligned}
p(\mu|D) &= \mathcal{N}(\mu|\mu_n, \lambda_n^{-1}) \\
\lambda_n &= \lambda_0 + n\lambda \\
\mu_n &= \frac{\overline{x}n\lambda + \mu_0\lambda_0}{\lambda_n}
\end{aligned}
$$

Precisions add

Convex combination

- Posterior predictive: Gaussian

$$
\begin{aligned}
p(x|D) &= \int p(x|\mu)p(\mu|D)d\mu \\
&= \int \mathcal{N}(x|\mu, \sigma^2)\mathcal{N}(\mu|\mu_n, \sigma_n^2)d\mu \\
&= \mathcal{N}(x|\mu_n, \sigma_n^2 + \sigma^2)
\end{aligned}
$$

Uncertainty about μ

noise

Real values

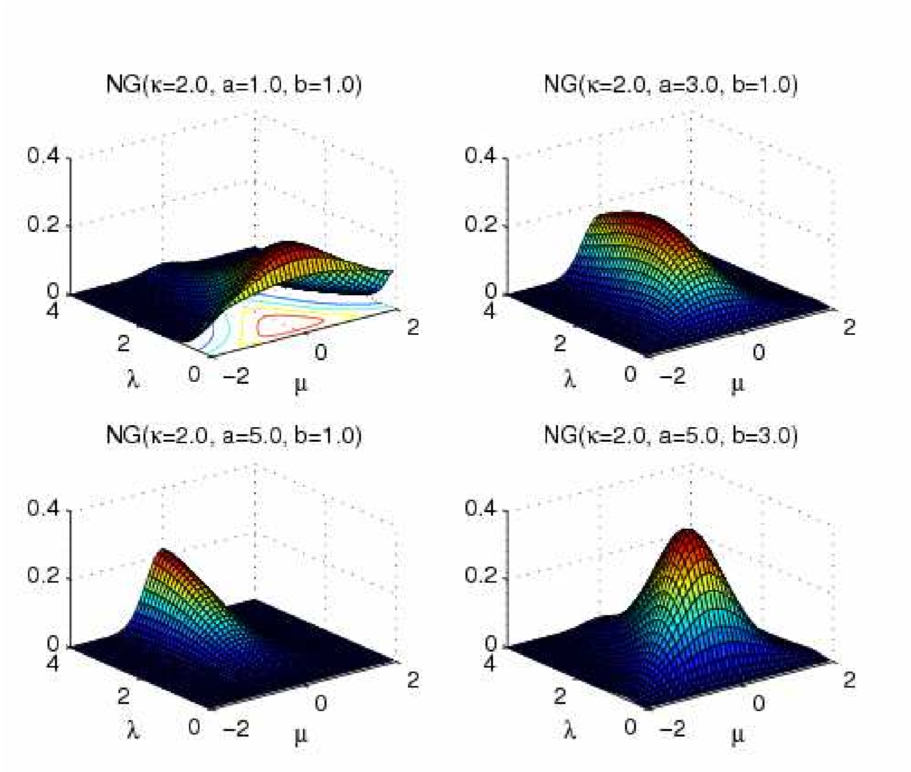
- Data: $x_i \in R$
- Hypothesis space: $\mu \in R$, $\lambda \in R^+$
- Likelihood: Gaussian

$$
\begin{aligned}
p(D|\mu) &= \frac{1}{(2\pi)^{n/2}} \lambda^{n/2} \exp\left(-\frac{\lambda}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right) \\
&= \frac{1}{(2\pi)^{n/2}} \lambda^{n/2} \exp\left(-\frac{\lambda}{2}\left[n(\mu-\overline{x})^2+\sum_{i=1}^{n}(x_i-\overline{x})^2\right]\right)
\end{aligned}
$$

Real values

- Prior: Normal Gamma

$$NG(\mu, \lambda | \mu_0, \kappa_0, \alpha_0, \beta_0) \stackrel{\text{def}}{=} \mathcal{N}(\mu | \mu_0, (\kappa_0 \lambda)^{-1}) Ga(\lambda | \alpha_0, \text{rate} = \beta_0)$$
$$\propto \lambda^{\frac{1}{2}} \exp(-\frac{\kappa_0 \lambda}{2} (\mu - \mu_0)^2) \lambda^{\alpha_0 - 1} e^{-\lambda \beta_0}$$



Real values

- Posterior: Normal Gamma

$$
\begin{aligned}
p(\mu, \lambda|D) &= NG(\mu, \lambda|\mu_n, \kappa_n, \alpha_n, \beta_n) \\
\mu_n &= \frac{\kappa_0\mu_0 + n\overline{x}}{\kappa_0 + n} \\
\kappa_n &= \kappa_0 + n \\
\alpha_n &= \alpha_0 + n/2 \\
\beta_n &= \beta_0 + \tfrac{1}{2}\sum_{i=1}^{n}(x_i - \overline{x})^2 + \frac{\kappa_0 n(\overline{x} - \mu_0)^2}{2(\kappa_0 + n)}
\end{aligned}
$$

- Posterior predictive: student T (long-tailed Gaussian)

$$
p(x|D) = t_{2\alpha_n}(x|\mu_n, \frac{\beta_n(\kappa_n + 1)}{\alpha_n\kappa_n})
$$

Posterior summaries

- Common to quote the posterior mean E[θ|D] as a point estimate
- 95% Credible interval $(\ell(D), u(D))$

$$p\left(\ell \leq \theta \leq u | D\right) \geq 0.95$$

- Can also summarize using samples from posterior

$$\theta^s \sim p(\theta|D)$$

MAP estimation

- We will often use the posterior mode as an approximation to the full posterior.

$$\begin{aligned}\hat{\theta}_{MAP} &= \arg\max_{\theta} p(\theta|D) \\ &= \arg\max_{\theta} \log p(D|\theta) + \log p(\theta)\end{aligned}$$

$$p(\theta|D) \approx \delta(\theta - \hat{\theta}_{MAP})$$

- This ignores uncertainty in our estimate, and will result in overconfident predictions.
- However, it is often computationally much cheaper than a fully Bayesian solution.
- If $p(\theta) \propto 1$ (uninformative prior), then MAP = MLE.

Topics

- Bayesian statistics
- → Information theory
- Decision theory

Information theory

- Entropy = min num bits to encode samples from p(x) using optimal code (and knowledge of p(x))

$$H(X) = -\sum_x p(x) \log_2 p(x)$$

- KL divergence = extra num bits to encode samples coming from p(x) using code based on q(x)

$$\begin{aligned} KL(p||q) &= \sum_x p(x) \log \frac{p(x)}{q(x)} \\ &= \underbrace{-\sum_x p(x) \log q(x)}_{\text{Cross entropy from p to q}} - H(p) \end{aligned}$$

Cross entropy from p to q

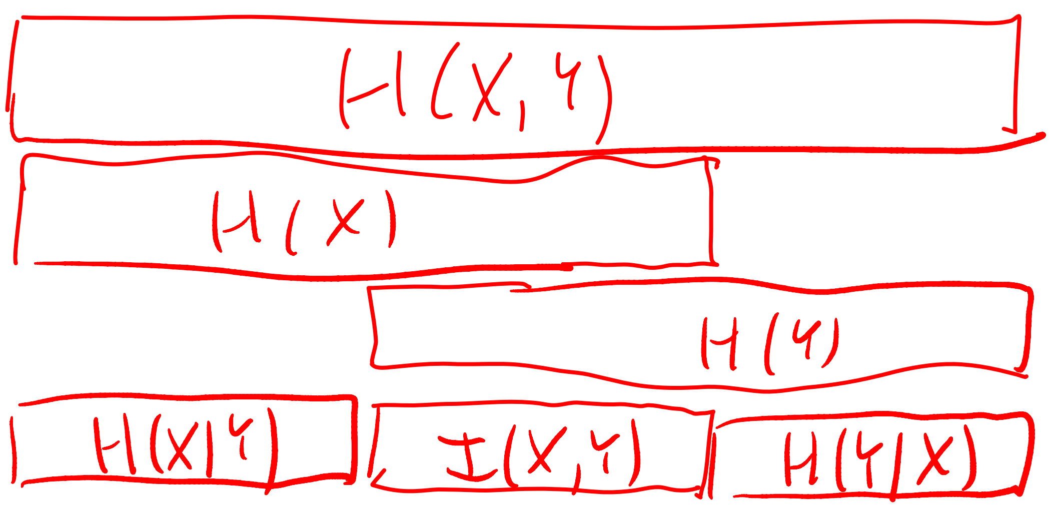
Mutual information

- I(X,Y) is how much our uncertainty about Y decreases when we observe X

$$
\begin{aligned}
I(X,Y) &\stackrel{\text{def}}{=} \sum_y \sum_x p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = KL(p(x,y)||p(x)p(y)) \\
&= -H(X,Y) + H(X) + H(Y) \\
&= H(X) - H(X|Y) = H(Y) - H(Y|X)
\end{aligned}
$$

- Hence

$$H(X,Y) = H(X|Y) + H(Y|X) + I(X,Y)$$



Mackay 9.1

Topics

- Bayesian statistics
- Information theory
- → Decision theory

Bayesian decision theory

- Pick action $\hat{\theta}(D)$ to minimize expected loss wrt current belief state p(θ|D)

$$\hat{\theta}(D) = \arg\min_{\hat{\theta}} EL(\theta, \hat{\theta}) = \arg\min_{\hat{\theta}} \int L(\theta, \hat{\theta}) p(\theta|D) d\theta$$

- Squared error (L2) loss

$$L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$$

Posterior mean

$$\frac{d}{d\hat{\theta}} EL(\theta, \hat{\theta}) = 0 \Rightarrow \hat{\theta}(D) = E[\theta|D]$$

- Zero-one loss

$$L(\theta, \hat{\theta}) = \delta(\theta - \hat{\theta})$$

Posterior mode

$$\frac{d}{d\hat{\theta}} EL(\theta, \hat{\theta}) = 0 \Rightarrow \hat{\theta}(D) = \arg\max_{\theta} p(\theta|D)$$

Decision theory: classifiers

- Given belief state p(y|x), pick action $\hat{y}(x)$ to minimize expected loss

$$\hat{y}(x) = \arg\min_{\hat{y}} EL(y, \hat{y}) = \arg\min_{\hat{y}} \sum_y L(y, \hat{y}) p(y|x)$$

state y

action $\hat{y}$

	1	2
1	True positive $\lambda_{11} = 0$	False positive λ_{12}
2	False negative λ_{21}	True negative $\lambda_{22} = 0$

$$\hat{y}(x) = 1 \text{ iff } \frac{p(Y = 1|x)}{p(Y = 2|x)} > \frac{\lambda_{12}}{\lambda_{21}}$$

Decision theory: model selection

- Given belief state p(m|D), pick model $\hat{m}(D)$ to minimize expected loss
- For 0-1 loss, pick most probable model

$$
\begin{aligned}
m^*(D) &= \arg\max_m p(m|D) \\
p(m|D) &= \frac{p(m)p(D|m)}{p(D)} \\
p(D) &= \sum_{m \in \mathcal{M}} p(m)p(D|m)
\end{aligned}
$$

Bayes factors

- To compare 2 models with equal priors, use the Bayes factor (c.f. likelihood ratio)

$$BF(i,j) \quad = \quad \frac{p(D|m_i)}{p(D|m_j)}$$

- The marginal likelihood p(D|m) is the probability that model m can generate D using parameters sampled from its prior

$$p(D|m) \quad = \quad \int p(D|\theta, m) p(\theta|m) d\theta$$

- This automatically penalizes complex models (Occam's razor)

Predicting the future

- Consider predicting $y=x_{n+1}$ given $x_{1:n}$.
- Use a loss function that measures your surprise

$$L(m, y) = -\log p(y|m)$$

- Pick m to minimize expected loss (risk)

$$R(m) = \int p(y|x_{1:n}) L(m, y) dy$$

$$= \int -p(y|x_{1:n}) \log p(y|x_{1:n}, m) = E_y f_m(y, x_{1:n})$$

- To minimize this cross entropy, we should pick the model whose predictions $p(y|x_{1:n},m)$ come closest to our predictions of the future, given by this Bayes model average

$$p(y|x_{1:n}) = \sum_{m \in \mathcal{M}} p(y|m, x_{1:n}) p(m|x_{1:n})$$

Cross validation

- If we don't think the true model is in our model class $\mathcal{M}$, we can approximate $p(y|x_{1:n})$ empirically.
- Leave one out cross validation (LOOCV) uses x_i as test and x_{-i} as training, and averages over i

$$E_y f_m(y, x_{1:n}) \quad \approx \quad \frac{1}{n}\sum_{i=1}^{n} f_m(x_i, x_{-i})$$

- We then pick the model m with the minimal empirical cross entropy loss.
- We can reduce the variance of this estimate using larger splits, eg 10-fold cross validation.