CS340: Machine Learning

Modelling discrete data with Bernoulli and multinomial distributions

Kevin Murphy

Modeling discrete data

- Some data is discrete/ symbolic, e.g., words, DNA sequences, etc.
- We want to build probabilistic models of discrete data $p(X|M)$ for use in classification, clustering, segmentation, novelty detection, etc.
- We will start with models (density functions) of a single **categorical** random variable $X \in \{1, \ldots, K\}$. (Categorical means the values are unordered, not low/ medium/ high).
- Today we will focus on $K = 2$ states, i.e., binary data.
- Later we will build models for multiple discrete random variables.

Bernoulli distribution

- Let $X \in \{0,1\}$ represent tails/ heads.
- Suppose $P(X=1)=\theta$. Then

$$P(x|\theta) = \text{Be}(X|\theta) = \theta^x(1-\theta)^{1-x}$$

- It is easy to show that

$$E[X] = \theta, \quad \text{Var}[X] = \theta(1-\theta)$$

- Given $D=(x_1,\ldots,x_N)$, the likelihood is

$$p(D|\theta) = \prod_{n=1}^{N} p(x_n|\theta) = \prod_{n=1}^{N} \theta^{x_n}(1-\theta)^{1-x_n} = \theta^{N_1}(1-\theta)^{N_0}$$

where $N_1 = \sum_n x_n$ is the number of heads and $N_0 = \sum_n (1-x_n)$ is the number of tails (sufficient statistics). Obviously $N = N_0 + N_1$.

Binomial distribution

- Let $X \in \{1, \ldots, N\}$ represent the number of heads in N trials. Then X has a binomial distribution

$$p(X|N) = \binom{N}{X} \theta^X (1-\theta)^{N-X}$$

where

$$\binom{N}{X} = \frac{N!}{(N-X)!X!}$$

is the number of ways to choose X items from N.

- We will rarely use this distribution.

Parameter estimation

- Suppose we have a coin with probability of heads θ. How do we estimate θ from a sequence of coin tosses $D = (X_1, \ldots, X_n)$, where $X_i \in \{0, 1\}$?
- One approach is to find a maximum likelihood estimate

$$\hat{\theta}_{ML} = \arg\max_{\theta} p(D|\theta)$$

- The Bayesian approach is to treat θ as a random variable and to use Bayes rule

$$p(\theta|D) = \frac{p(\theta)p(D|\theta)}{\int_{\theta'} p(\theta', D)}$$

  and then to return the posterior mean or mode.
- We will discuss both methods below.

MLE (maximum likelihood estimate) for Bernoulli

- Given $D = (x_1, \ldots, x_N)$, the likelihood is

$$p(D|\theta) = \theta^{N_1}(1-\theta)^{N_0}$$

- The log-likelihood is

$$L(\theta) \; = \; \log p(D|\theta) = N_1 \log \theta + N_0 \log(1-\theta)$$

- Solving for $\frac{dL}{d\theta} = 0$ yields

$$\theta_{ML} = \frac{N_1}{N_1 + N_0} = \frac{N_1}{N}$$

Problems with the MLE

- Suppose we have seen $N_1 = 0$ heads out of $N = 3$ trials. Then we predict that heads are impossible!
$$\theta_{ML} = \frac{N_1}{N} = \frac{0}{3} = 0$$
- This is an example of the *sparse data problem*: if we fail to see something in the training set (e.g., an unknown word), we predict that it can never happen in the future.
- We will now see how to solve this pathology using Bayesian estimation.

Bayesian parameter estimation

- The Bayesian approach is to treat θ as a random variable and to use Bayes rule
$$p(\theta|D) = \frac{p(\theta)p(D|\theta)}{\int_{\theta'} p(\theta', D)}$$
- We need to specify a prior $p(\theta)$. This reflects our subjective beliefs about what possible values of θ are plausible, before we have seen any data.
- We will discuss various "objective" priors below.

The beta distribution

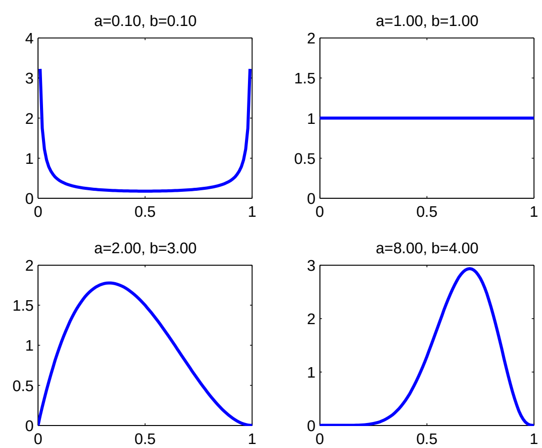
We will assume the prior distribution is a beta distribution,

$$p(\theta) = Be(\theta|\alpha_1, \alpha_0) \propto [\theta^{\alpha_1 - 1}(1-\theta)^{\alpha_0 - 1}]$$

This is also written as $\theta \sim Be(\alpha_1, \alpha_0)$ where α_0, α_1 are called **hyper-parameters**, since they are parameters of the prior. This distribution satisfies

$$E\theta = \frac{\alpha_1}{\alpha_0 + \alpha_1}$$

$$\text{mode } \theta = \frac{\alpha_1 - 1}{\alpha_0 + \alpha_1 - 2}$$



Conjugate priors

- A prior $p(\theta)$ is called conjugate if, when multiplied by the likelihood $p(D|\theta)$, the resulting posterior is in the same parametric family as the prior. (Closed under Bayesian updating.)
- The Beta prior is conjugate to the Bernoulli likelihood

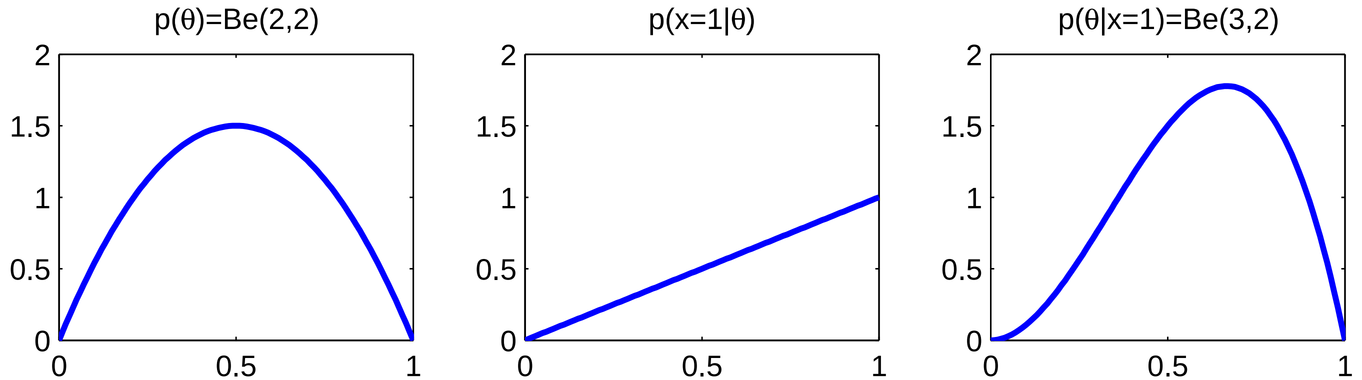
$$\begin{aligned} P(\theta|D) &\propto P(D|\theta)P(\theta) = p(D|\theta)Be(\theta|\alpha_1, \alpha_0) \\ &\propto [\theta^{N_1}(1-\theta)^{N_0}][\theta^{\alpha_1 - 1}(1-\theta)^{\alpha_0 - 1}] \\ &= \theta^{N_1 + \alpha_1 - 1}(1-\theta)^{N_0 + \alpha_0 - 1} \\ &\propto Be(\theta|\alpha_1 + N_1, \alpha_0 + N_0) \end{aligned}$$

- e.g., start with $Be(\theta|2,2)$ and observe $x = 1$ to get $Be(\theta|3,2)$, so the mean shifts from $E[\theta] = 2/4$ to $E[\theta|D] = 3/5$.
- We see that the hyperparameters α_1, α_0 act like "pseudo counts", and correspond to the number of "virtual" heads/tails.
- $\alpha = \alpha_0 + \alpha_1$ is called the effective sample size (strength) of the prior, since it plays a role analogous to $N = N_0 + N_1$.

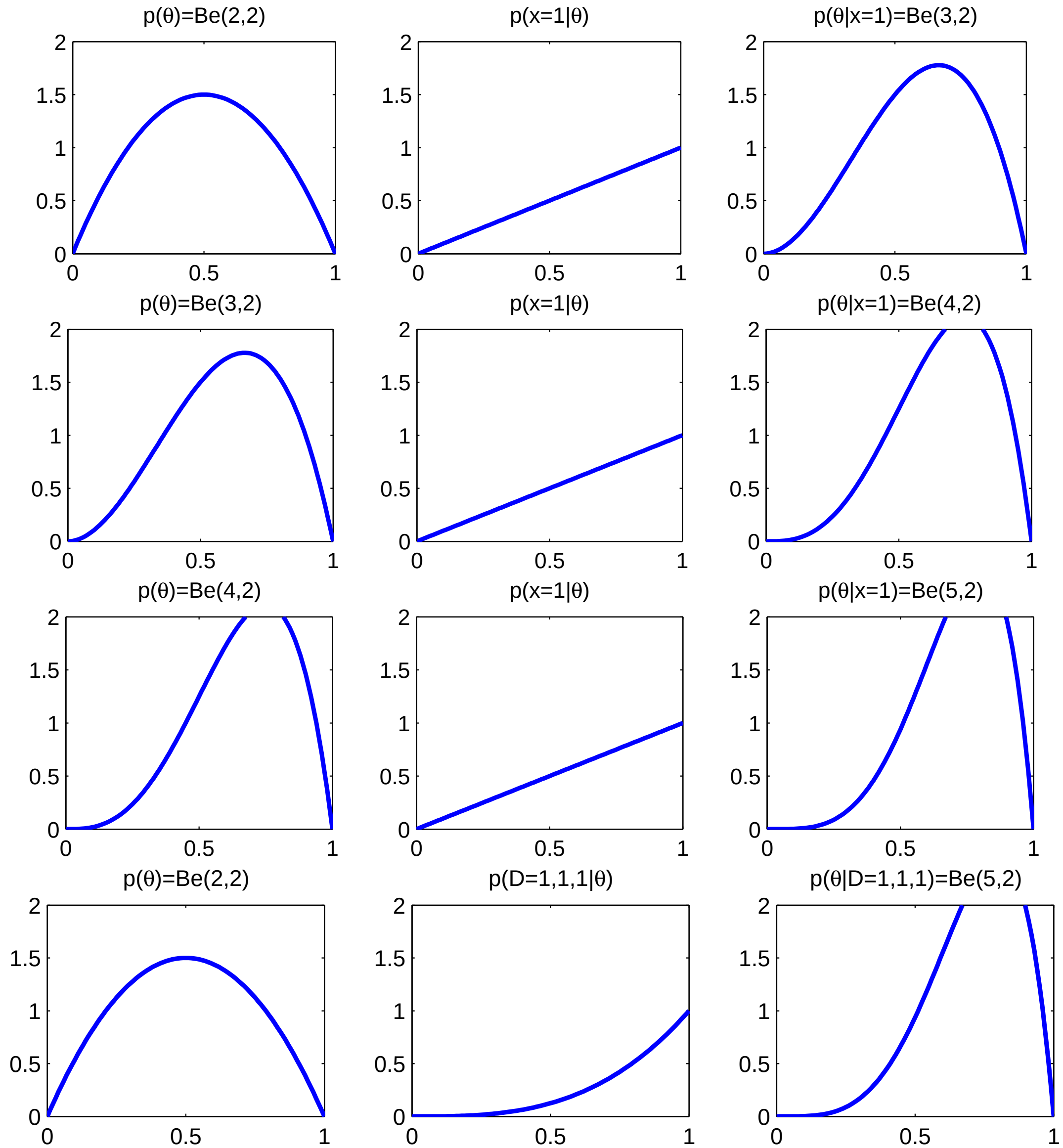
Bayesian updating in pictures

- Start with $Be(\theta|\alpha_0 = 2, \alpha_1 = 2)$ and observe $x = 1$, so the posterior is $Be(\theta|\alpha_0 = 3, \alpha_1 = 2)$.

```
thetas = 0:0.01:1;
alpha1 = 2; alpha0 = 2; N1=1; N0=0; N = N1+N0;
prior = betapdf(thetas, alpha1, alpha1);
lik =  thetas.^N1 .* (1-thetas).^N0;
post = betapdf(thetas, alpha1+N1, alpha0+N0);
subplot(1,3,1);plot(thetas, prior);
subplot(1,3,2);plot(thetas, lik);
subplot(1,3,3);plot(thetas, post);
```



Sequential Bayesian updating



Sequential Bayesian updating

- Start with $Be(\theta|\alpha_1, \alpha_0)$ and observe N_0, N_1 to get $Be(\theta|\alpha_1 + N_1, \alpha_0 + N_0)$.
- Treat the posterior as a new prior: define $\alpha_0' = \alpha_0 + N_0$, $\alpha_1' = \alpha_1 + N_1$, so $p(\theta|N_0, N_1) = Be(\theta|\alpha_1', \alpha_0')$.
- Now see a new set of data, N_0', N_1' to get get the new posterior

$$\begin{aligned} p(\theta|N_0, N_1, N_0', N_1') &= Be(\theta|\alpha_1' + N_1', \alpha_0' + N_0') \\ &= Be(\theta|\alpha_1 + N_1 + N_1', \alpha_0 + N_0 + N_0') \end{aligned}$$

- This is equivalent to combining the two data sets into one big data set with counts $N_0 + N_0'$ and $N_1 + N_1'$.
- The advantage of sequential updating is that you can learn online, and don't need to store the data.

Point estimates

- $p(\theta|D)$ is the full posterior distribution. Sometimes we want to collapse this to a single point. It is common to pick the posterior mean or posterior mode.
- If $\theta \sim Be(\alpha_1, \alpha_0)$, then $E\theta = \frac{\alpha_1}{\alpha}$, mode $\theta = \frac{\alpha_1 - 1}{\alpha - 2}$.
- Hence the MAP (maximum a posterior) estimate is

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(D|\theta)p(\theta) = \frac{\alpha_1 + N_1 - 1}{\alpha + N - 2}$$

- The posterior mean is

$$\hat{\theta}_{mean} = \frac{\alpha_1 + N_1}{\alpha + N}$$

- The maximum likelihood estimate is

$$\hat{\theta}_{MLE} = \frac{N_1}{N}$$

Posterior predictive distribution

- The posterior predictive distribution is

$$\begin{aligned} p(X=1|D) &= \int_0^1 p(X=1|\theta)p(\theta|D)d\theta \\ &= \int_0^1 \theta \;\; p(\theta|D)d\theta = E[\theta|D] \\ &= \frac{N_1+\alpha_1}{N_1+N_0+\alpha_1+\alpha_0} = \frac{N_1+\alpha_1}{N+\alpha} \end{aligned}$$

- With a uniform prior $\alpha_0 = \alpha_1 = 1$, we get Laplace's rule of succession

$$p(X=1|N_1,N_0) = \frac{N_1+1}{N_1+N_0+2}$$

- eg. if we see $D = 1, 1, 1, \ldots$, our predicted probability of heads steadily increases: $\frac{1}{2}$, $\frac{2}{3}$, $\frac{3}{4}$, ...

Plug-in estimates

- Rather than integrating over the posterior, we can pick a single point estimate of θ and make predictions using that.

$$
\begin{aligned}
p(X = 1|D, \hat{\theta}_{ML}) &= \frac{N_1}{N} \\
p(X = 1|D, \hat{\theta}_{mean}) &= \frac{N_1 + \alpha_1}{N + \alpha} \\
p(X = 1|D, \hat{\theta}_{MAP}) &= \frac{N_1 + \alpha_1 - 1}{N + \alpha - 2}
\end{aligned}
$$

- *In this case* the full posterior predictive density $p(X = 1|D)$ is the same as the plug-in estimate using the posterior mean parameter $p(X = 1|D, \hat{\theta}_{mean})$.

Posterior mean

- The posterior mean is a convex combination of the prior mean $\alpha'_1 = \alpha_1/\alpha$ and the MLE N_1/N:

$$\begin{aligned}\hat{\theta}_{mean} &= \frac{\alpha_1 + N_1}{\alpha + N} \\ &= \frac{\alpha'_1 \alpha}{\alpha + N} + \frac{N}{\alpha + N}\frac{N_1}{N} \\ &= \lambda \alpha'_1 + (1-\lambda)\frac{N_1}{N}\end{aligned}$$

where

$$\lambda = \frac{\alpha}{N + \alpha}$$

is the prior weight relative to the total weight.

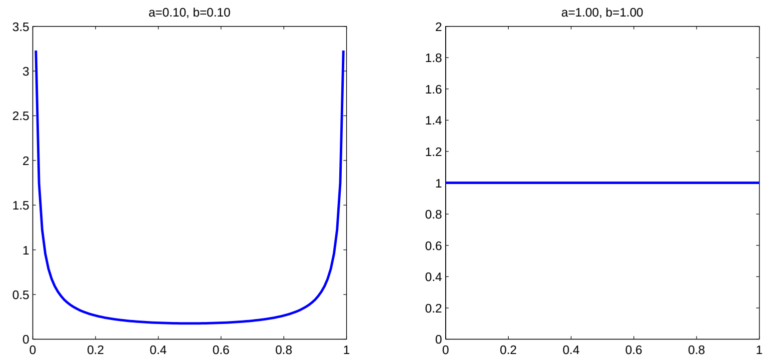
- (We will derive a similar result later for Gaussians.)

Effect of prior strength

- Suppose we weakly believe in a fair coin, $p(\theta) = Be(1, 1)$.
- If $N_1 = 3, N_0 = 7$ then $p(\theta|D) = Be(4, 8)$ so $E[\theta|D] = 4/12 = 0.33$.
- Suppose we strongly believe in a fair coin, $p(\theta) = Be(10, 10)$.
- If $N_1 = 3, N_0 = 7$ then $p(\theta|D) = Be(13, 17)$ so $E[\theta|D] = 13/30 = 0.43$.
- With a strong prior, we need a lot of data to move away from our initial beliefs.

Uninformative/ objective/ reference prior

- If $\alpha_0 = \alpha_1 = 1$, then $Be(\theta|\alpha_1, \alpha_0)$ is uniform, which seems like an uninformative prior.



- But since the posterior predictive is

$$p(X = 1|N_1, N_0) = \frac{N_1 + \alpha_1}{N + \alpha}$$

$\alpha_1 = \alpha_0 = 0$ is a better definition of uninformative, since then the posterior mean is the MLE.

- Note that as $\alpha_0, \alpha_1 \rightarrow 0$, the prior becomes bimodal.
- This shows that a uniform prior is not always uninformative.

From coins to dice: multinomial distribution

- Let $X \in \{1, \ldots, K\}$ have distribution
$$p(X = k|\theta) = \theta_k = \theta_1^{I(X=1)} \theta_2^{I(X=2)} \cdots \theta_K^{I(X=k)}$$
This is called a multinomial distribution. We require $0 \leq \theta_k \leq 1$ and $\sum_{k=1}^{K} \theta_k = 1$.
- $I(e) = 1$ if event e is true, and $I(e) = 0$ otherwise (the indicator function).
- e.g., a fair dice has $\theta_k = 1/6$ for $k = 1 : 6$.
- Sometimes instead of writing $X = k$ we will use a one-of-K encoding. Specifically, $[x] \in \{0, 1\}^K$ with the k'th bit on means $X = k$. eg. if $x = 3$ and $K = 6$, then $[x] = (0, 0, 1, 0, 0, 0)$.

Maximum likelihood estimation

- Suppose we observe N iid die rolls (K-sided): D=3,1,6,2,...
- The log likelihood of the data is given by

$$\begin{aligned}\ell(\theta; D) &= \log p(D|\theta) = \log \prod_m p(x_m|\theta) \\ &= \sum_m \log \prod_k \theta_k^{I(x^m=k)} \\ &= \sum_m \sum_k I(x^m = k) \log \theta_k = \sum_k N_k \log \theta_k\end{aligned}$$

- The sufficient statistics are the counts $N_k = \sum_m I(X_m = k)$,
- We need to maximize this subject to the constraint $\sum_k \theta_k = 1$, so we use a Lagrange multiplier.

Maximum likelihood estimation

- Constrained cost function:

$$\tilde{l} = \sum_k N_k \log \theta_k + \lambda \left(1 - \sum_k \theta_k \right)$$

- Take derivatives wrt θ_k:

$$\begin{aligned}
\frac{\partial \tilde{l}}{\partial \theta_k} &= \frac{N_k}{\theta_k} - \lambda = 0 \\
N_k &= \lambda \theta_k \\
\sum_k N_k &= N = \lambda \sum_k \theta_k = \lambda \\
\hat{\theta}_k &= \frac{N_k}{N}
\end{aligned}$$

- $\hat{\theta}_k$ is the fraction of times k occurs.

MLE Example

- Suppose $K = 6$ and we see $D = (1, 6, 1, 2)$ so $N = 4$. Then
$$\hat{\theta} = (2/4, 1/4, 0/4, 0/4, 0/4, 1/4)$$

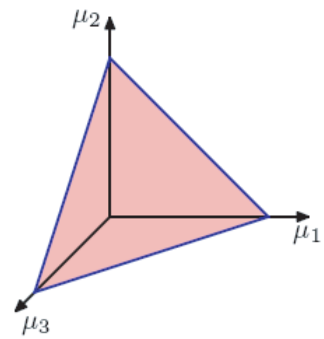
Bayesian estimation

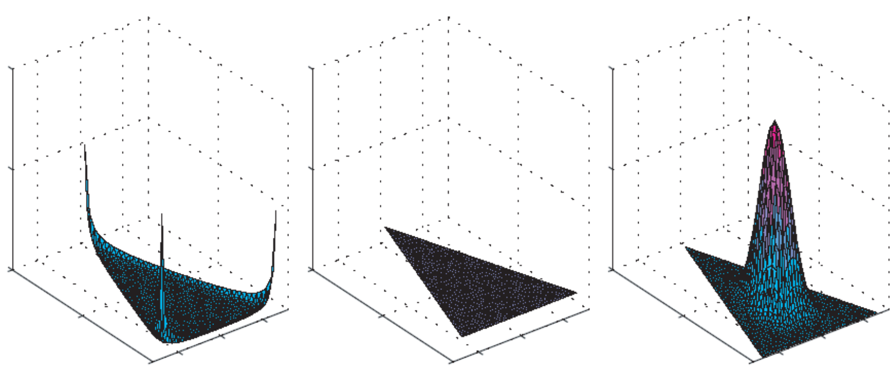
- We will now consider Bayesian estimates $p(\theta|D)$.
- We just replace the bernoulli likelihood with a multinomial likelihood, and replace the beta prior with a Dirichlet prior.

Dirichlet priors

A Dirichlet prior generalizes the beta from binary variables to K-ary variables.

$$p(\theta|\alpha) = \mathcal{D}(\theta|\alpha) \propto \theta_1^{\alpha_1 - 1} \cdot \theta_2^{\alpha_2 - 1} \cdots \theta_K^{\alpha_K - 1}$$





Properties of the Dirichlet distribution

- If $\theta \sim Dir(\theta|\alpha_1, \ldots, \alpha_K)$, then

$$E[\theta_k] = \frac{\alpha_k}{\alpha}$$
$$\text{mode}[\theta_k] = \frac{\alpha_k - 1}{\alpha - K}$$

where $\alpha \stackrel{\text{def}}{=} \sum_{k=1}^{K} \alpha_k$ is the total strength of the prior.

Dirichlet-multinomial model

By analogy to the Beta-bernoulli case, we can just write down the likelihood, prior, posterior and predictive as follows

$$
\begin{aligned}
P(\vec{N}|\vec{\theta}) &= \prod_{i=1}^{K} \theta_i^{N_i} \\
p(\theta|\alpha) &= \mathcal{D}(\theta|\alpha) \propto \theta_1^{\alpha_1 - 1} \cdot \theta_2^{\alpha_2 - 1} \cdots \theta_K^{\alpha_K - 1} \\
p(\theta|\vec{N}, \vec{\alpha}) &= \mathcal{D}(\alpha_1 + N_1, \ldots, \alpha_K + N_K) \\
p(X = k|D) &= E[\theta_k|D] = \frac{N_k + \alpha_k}{N + \alpha}
\end{aligned}
$$