CPSC 440/540: Machine Learning

Bayesian Learning

Winter 2023

Last Time: Monte Carlo Methods

- Monte Carlo approximates expectation of random functions:

$$\mathbb{E}[g(X)] = \sum_{x \in \mathcal{X}} g(x)p(x) \qquad \mathbb{E}[g(X)] = \int_{x \in \mathcal{X}} g(x)\,p(x)\,dx$$

pmf of discrete variable X

pdf of continuous X

  - Approximation is average of function *g* applied to samples from *p*:

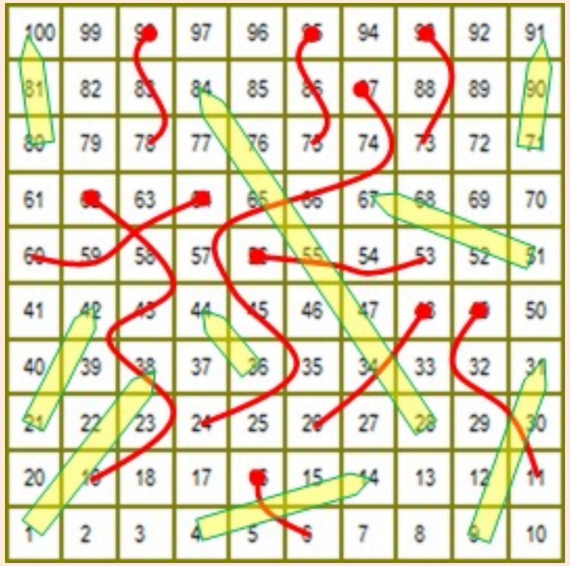
$$\mathbb{E}[g(X)] \approx \frac{1}{n}\sum_{i=1}^{n} g(x^i)$$

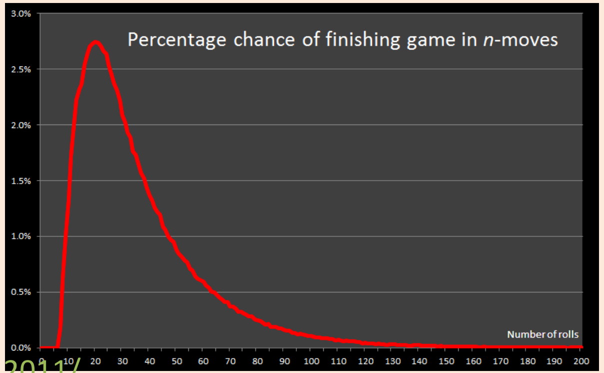
- Can approximate a wide variety of quantities by changing *g*:
  - Mean: g(x) = x.
  - Probability of event 'A': g(x) = 𝟙["A happened"].
  - CDF: g(x) = 𝟙[x ≤ c].
- This is useful when:
  - You know how to sample from p(x).
  - You do not know how to efficiently compute $\mathbb{E}$[g(x)].
  - Are patient and/or don't care about being precise, because it converges slowly.

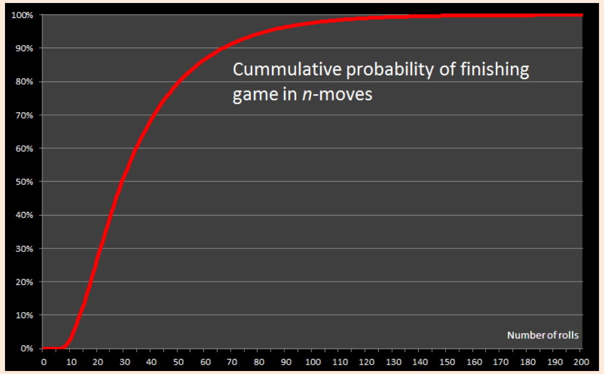
bonus!

Monte Carlo for Snakes and Ladders

- Consider the children's game "Snakes and Ladders":
  - Start on '1', roll die, move marker, go up/down on ladder/snake, end at 100.
  - No decisions, so you can simulate the game.
- How many turns does it take for this game to end?
  - Simulate game many times, count number of turns.
  - Compute average number of turns.
- Probability and cumulative probability:







https://www.datagenetics.com/blog/november12011/

Conditional Probabilities with Monte Carlo

- We often want to compute conditional probabilities.
  - "What is the probability that the game will go more than 100 turns, if it already went 50 turns?"

- A Monte Carlo approach for estimating p(*A* | *B*):
  - Generate a large number of samples.
  - Use Monte Carlo estimate of p(*A*, *B*) and p(*B*) to approximate conditional:

$$p(A \mid B) = \frac{p(A, B)}{p(B)} \approx \frac{\sum_{i=1}^{n} \mathbb{1}[\text{"A and B happened"}]}{\sum_{i=1}^{n} \mathbb{1}[\text{"B happened"}]}$$

    - Frequency of the first event, in samples consistent with the second event.
      - This is the MLE for a binary variable that is 1 when *A* happens, conditioned on *B* happening.

- This is a special case of rejection sampling (general case later).
  - Unfortunately, if *B* is rare then most samples are "rejected" (ignored).
  - The conditional probability demo here has a good visualization of this.

Next Topic: MLE and MAP for Categorical

MLE for Categorical Distribution

- Now we will consider how to train a categorical model ("learning").
  - Goal is to go from samples to an estimate of parameters $\theta_1, \theta_2, \dots, \theta_k$:

X =

Party?
LIB
CPC
NDP
LIB
GRN

p(x = LIB) = 0.34, p(x=NDP) = 0.34,
p(x = CPC) = 0.27, p(x=GRN) = 0.03,
p(x = PPC) = 0.02.

- As before we will first consider maximum likelihood estimation:

$$\hat{\Theta} \in \underset{\Theta}{\operatorname{argmax}}\{p(x^1, x^2, \dots, x^n \mid \Theta)\}$$

$\{\theta_1, \theta_2, \dots, \theta_k\}$

  - In this case the MLE is given by $\theta_c = \frac{n_c}{n}$ (n_c is number 'c' examples).
    - If "34% of your samples are LIB, your guess for θ_{LIB}=0.34".
    - As with Bernoulli, the derivation of the MLE is not as a simple as the result.

bonus!

Derivation of MLE (that does not work)

- Last time we showed that the likelihood has the form:

$$p(X \mid \Theta) = \theta_1^{n_1} \theta_2^{n_2} \cdots \theta_k^{n_k}$$

$x^1, x^2, \dots, x^n$ $\qquad$ $\theta_1, \theta_2, \dots, \theta_k$

- Let's take the log:

$$\log p(X \mid \Theta) = n_1 \log \theta_1 + n_2 \log \theta_2 + \cdots n_k \log \theta_k$$

- Take the derivative for a particular θ_c:

$$\nabla_{\theta_c} \log p(X \mid \Theta) = \frac{n_c}{\theta_c}$$

- Set derivative equal to zero:

$$0 = \frac{n_c}{\theta_c}$$

- …huh?

Derivation of MLE: Handling "Sum to 1"

- "Set derivative of log-likelihood equal to 0" does not work.
  - Because of constraint that the θ_c must sum to 1, derivative is not zero at MLE.

- Approaches used in textbooks to enforce constraints:
  - Use "Lagrange multipliers" and find stationary point of "Lagrangian".
  - Define $\theta_k = 1 - \sum_{c=1}^{k-1} \theta_c$ to make it unconstrained.
  - See StackExchange thread here.

- We will take a different approach to make it unconstrained:
  1. Use a unnormalized parameterization $\tilde{\theta}_c$ that doesn't have constraints.
  2. Compute the MLE for the $\tilde{\theta}_c$ by setting log-likelihood derivative to zero.
  3. Convert from the $\tilde{\theta}_c$ parameters to our usual θ_c parameters by normalizing.

Unconstrained Parameterization

- Consider categorical distribution with unnormalized parameters:

$$p(x=c \mid \tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_k) \propto \tilde{\theta}_c$$

  - To give non-negative probabilities, we require that $\bar{\theta}_c \geq 0$ for all 'c'.
- The normalized probability can then be written:

$$p(x=c \mid \tilde{\Theta}) = \frac{\tilde{\theta}_c}{\sum_{c'=1}^{k} \tilde{\theta}_{c'}} = \frac{\tilde{\theta}_c}{Z}$$

$Z = \sum_{c'=1}^{k} \tilde{\theta}_{c'}$ is called the "normalizing constant"

It is constant in terms of 'x' but a function of $\tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_k$

  - The "normalizing constant" makes the probability sum to 1 across *c* values.
    - So we do not need to an explicit "sum to 1" constraint.
  - We convert from unnormalized to normalized by dividing by *Z*: $\theta_c = \frac{\tilde{\theta}_c}{Z}$.

Derivation of MLE (that does work)

- Using the unnormalized parameters in the likelihood gives:

$$p(X \mid \Theta) = \left(\frac{\tilde{\theta}_1}{Z}\right)^{n_1}\left(\frac{\tilde{\theta}_2}{Z}\right)^{n_2}\cdots\left(\frac{\tilde{\theta}_k}{Z}\right)^{n_k} = \frac{\tilde{\theta}_1^{n_1}\tilde{\theta}_2^{n_2}\cdots\tilde{\theta}_k^{n_k}}{Z^n}$$

- Let's take the log: $\log p(X \mid \Theta) = n_1 \log(\tilde{\theta}_1) + n_2 \log(\tilde{\theta}_2) + \cdots + n_k \log(\tilde{\theta}_k) - n \log Z$

- Take the derivative for a particular θ_c: $\nabla_{\tilde{\theta}_c} p(X \mid \Theta) = \frac{n_c}{\tilde{\theta}_c} - \frac{n}{Z}$

- Set derivative equal to zero: $0 = \frac{n_c}{\tilde{\theta}_c} - \frac{n}{Z}$

- Solve for $\tilde{\theta}_c$: $\frac{\tilde{\theta}_c}{Z} = \frac{n_c}{n}$ $\longrightarrow$ Convert to normalized: $\theta_c = \frac{n_c}{n}$

(and possible to show this maximizes likelihood)

$\log Z = \log\left(\sum_{c=1}^{k} \tilde{\theta}_c\right)$

$\nabla_{\tilde{\theta}_c} \log Z = \frac{1}{\sum_{c=1}^{k} \tilde{\theta}_c} = \frac{1}{Z}$

MAP Estimation and Dirichlet Prior

- As before, we may prefer to use a MAP estimate over the MLE.
  - Often becomes more important as *k* grows.
    - More parameters to [over]fit.

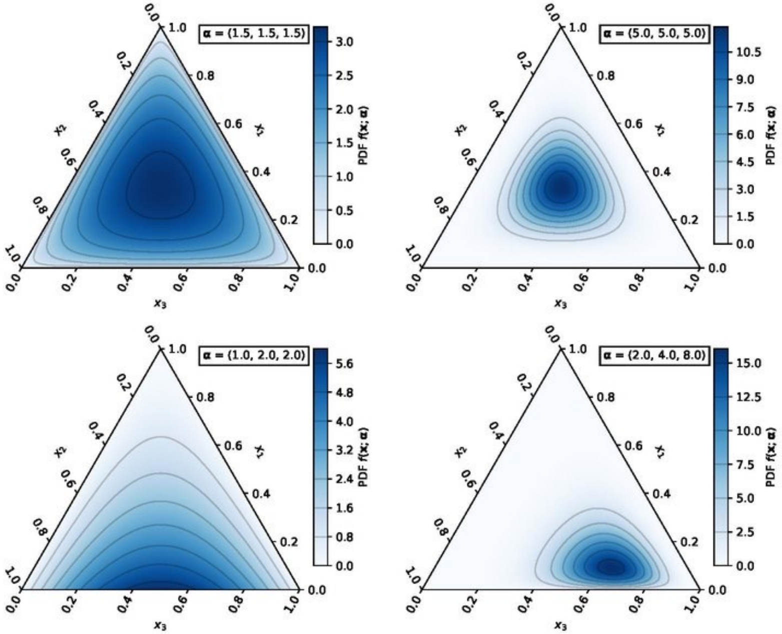
- Most common prior for categorical is the Dirichlet distribution:

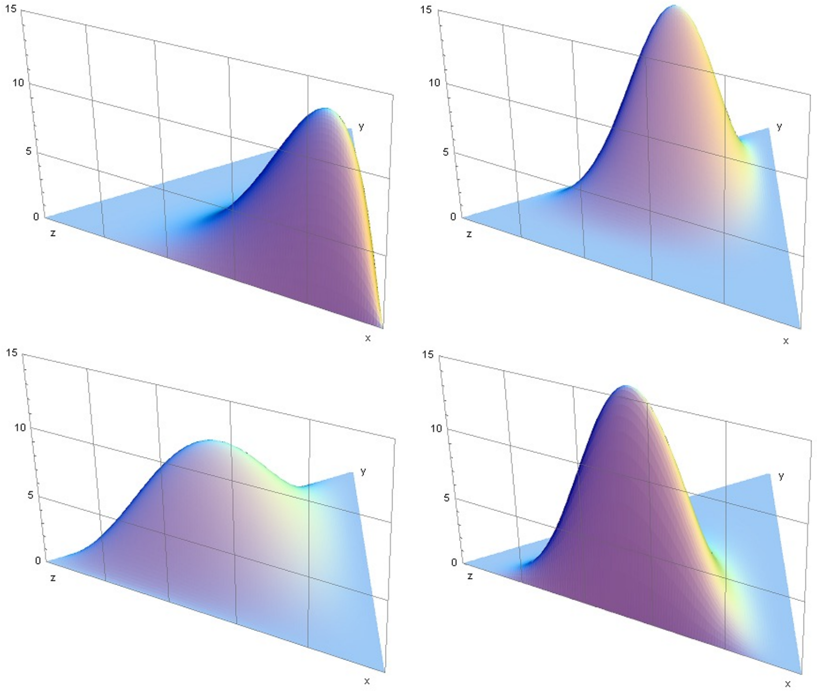
$$p(\theta_1, \theta_2, \ldots, \theta_k \mid \alpha_1, \alpha_2, \ldots, \alpha_k) \propto \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \cdots \theta_k^{\alpha_k - 1}$$

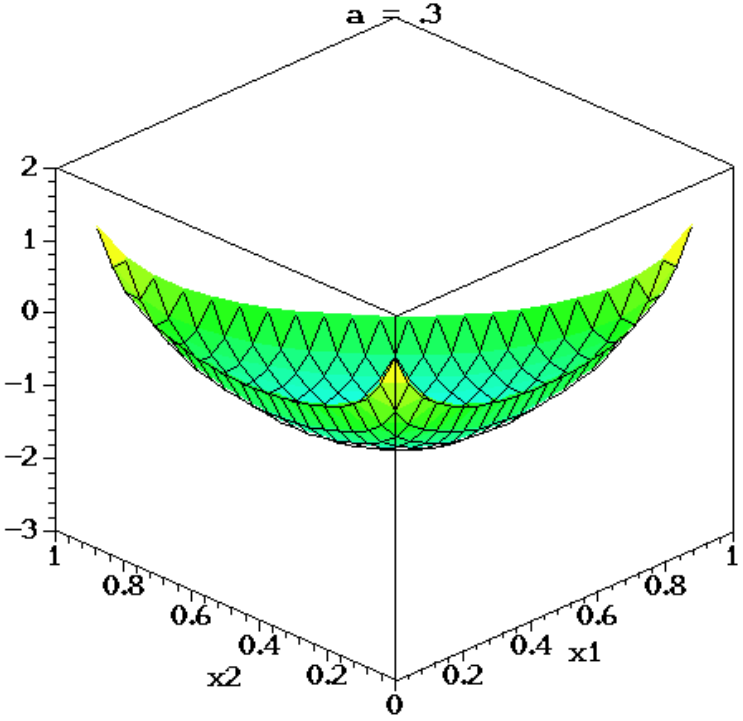
  - Generalization of the beta distribution to *k* classes (requires $\alpha_c > 0$).
- This is a distribution over Θ values:
  - Since the Θ parameterize probabilities,
    Dirichlet is a probability distribution over possible probability distributions.

Dirichlet Distribution

- Wikipedia's visualizations of Dirichlet distribution for k=3:







all α_c equal

- Can bias towards various types of probabilities.

https://en.wikipedia.org/wiki/Dirichlet_distribution
https://commons.wikimedia.org/wiki/Category:Dirichlet_distribution

MAP Estimation and Dirichlet Prior

- The MAP for a categorical with Dirichlet prior is given by:

$$\hat{\Theta}_c = \frac{n_c + \alpha_c - 1}{\sum_{c'=1}^{k}[n_{c'} + \alpha_{c'} - 1]}$$

  – Derivation is similar to the MLE derivation.

- Dirichlet has *k* hyper-parameters α_c.
  – We often set $\alpha_c = \alpha$ for some constant α (reduces to 1 hyper-parameter).
  – This simplifies the MLE to:

$$\hat{\Theta}_c = \frac{n_c + \alpha - 1}{\sum_{c'=1}^{k} n_{c'} + k(\alpha - 1)}$$

  – With $\alpha = 2$, we get Laplace smoothing ("add 1 to count of each class").

Posterior for Categorical Likelihood + Dirichlet Prior

- People use the Dirichlet because posterior has a simple form:

$$p(\Theta \mid X, \alpha) \propto p(X \mid \Theta)\, p(\Theta \mid \alpha) \propto \theta_1^{n_1} \theta_2^{n_2} \cdots \theta_k^{n_k}\, \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \cdots \theta_k^{\alpha_k - 1}$$

$\{\theta_1, \theta_2, \ldots, \theta_k\}$ $(\alpha_1, \alpha_2, \ldots, \alpha_k)$

Assuming data is independent of parameters given hyper-parameters

$$= \theta_1^{(n_1+\alpha_1)-1} \theta_2^{(n_2+\alpha_2)-1} \cdots \theta_k^{(n_k+\alpha_k)-1}$$

$$= \theta_1^{\tilde{\alpha}_1 - 1} \theta_2^{\tilde{\alpha}_2 - 1} \cdots \theta_k^{\tilde{\alpha}_k - 1}$$

  – This is another Dirichlet distribution with "updated" parameters $\tilde{\alpha}_c$.
    - Where $\tilde{\alpha}_c = n_c + \alpha_c$.
    - Again, make sure you understand why we can recognize this as a Dirichlet.
      – The normalizing constant must be the normalizing constant for the Dirichlet.

$$Z = \int_0^1 \int_0^1 \cdots \int_0^1 \theta_1^{\tilde{\alpha}_1 - 1} \theta_2^{\tilde{\alpha}_2 - 1} \cdots \theta_k^{\tilde{\alpha}_k - 1}\, d\theta_1\, d\theta_2 \cdots d\theta_k$$

Conjugate Priors

- We have now some examples of a convenient phenomenon:
  - If we put a beta prior on a Bernoulli likelihood, posterior is beta.
    - Same happens if you put beta prior on binomial/geometric: posterior is beta.
  - If we put a Dirichlet prior on a categorical likelihood, posterior is Dirichlet.

- In these situations, we say the prior is conjugate to the likelihood.
  - With conjugate priors, the prior and posterior come from the same "family".

$$x \sim D(\theta), \quad \theta \sim P(\lambda) \quad \Rightarrow \quad \theta \mid x \sim P(\lambda')$$

this means "has the probability distribution of"

    - The posterior will look like the prior with "updated" parameters.

- Many computations become easier when we use conjugate priors.
  - Because we have an explicit formula for the posterior distribution.
  - But not all distributions have conjugate priors.

Next Topic: Bayesian Learning

Problems with MAP

- With good hyper-parameters, MAP usually outperforms MLE.

- But MAP is still weird.
  - Recall that we said that decoding the mode can do weird things.
    - The value with highest probability/PDF may not represent "typical" behavior.
  - MAP is *maximum a posteriori*, the posterior mode.

- MAP is fine if you want to find parameters with highest probability, but in ML usually the goal is to make predictions (or decisions).
  - Our ultimate goal is not just to find the best parameters.

- You can show that MAP is a sub-optimal way to make predictions.

Example: "Two Heads" with "Fair vs. Unfair" Prior

- Suppose you have a Bernoulli variable and the following prior:
  - p(θ = 0.5) = 0.5 and p(θ = 1) = 0.5.
    - You think coin has 50% chance of being fair, 50% chance of "always landing head".
- The first two coin flips are "head".
  - $x^1 = 1$, $x^2 = 1$.
- What is the probability that the third flip will be a "head"?
  - MAP approach:

1. Find $\hat{\theta} \in \text{argmax}\{p(\theta | X)\} \equiv \text{argmax}\{p(X|\theta)p(\theta)\}$
2. Compute $p(x^3 = 1 | \hat{\theta} = 1) = 1$

$\theta = \frac{1}{2}$: $(\frac{1}{2})(\frac{1}{2})(\frac{1}{2}) = \frac{1}{8}$

$\theta = 1$: $(1)(1)(\frac{1}{2}) = \frac{1}{2}$

Since $\frac{1}{2} > \frac{1}{8}$, set $\hat{\theta} = 1$

  - MAP predicts 100% chance of head.
    - But the MAP "decoding" of the parameters is over-confident.
      - There was a 1/4 chance of seeing two heads from the fair coin.

Example: "Two Heads" with "Fair vs. Unfair" Prior

- Can compute correct probability using marginalization rule over θ:

$$\underbrace{p(x^3=1 \mid X)}_{\text{the probability we want}} \overset{\text{marg. rule}}{=} \sum_{\theta \in \{0.5,1\}} p(x^3=1, \theta \mid X) \overset{\text{product rule}}{=} \sum_{\theta \in \{0.5,1\}} \underbrace{p(x^3=1 \mid \theta, X)}_{\text{prediction given } \theta} \underbrace{p(\theta \mid X)}_{\text{posterior}}$$

- The correct probability weights possible predictions by posterior.
  - Assume x^3 is independent of X once we know θ: $p(x^3=1 \mid \theta, X) = p(x^3=1 \mid \theta)$
  - Use Bayes rule to compute posterior and get final answer:

$$p(\theta \mid X) = \frac{p(X \mid \theta)p(\theta)}{\sum_{\theta'} p(X \mid \theta')p(\theta')}$$

$\theta = 1/2$: $\frac{1/8}{1/8 + 1/2} = \frac{1}{5}$

$\theta = 1$: $\frac{1/2}{1/8 + 1/2} = \frac{4}{5}$

plug in

$$p(x^3=1 \mid X) = \underbrace{(1/2)\cdot\left(\frac{1}{5}\right)}_{\text{probability from "fair" case}} + \underbrace{(1)\left(\frac{4}{5}\right)}_{\text{probability from "unfair" case}} = \frac{9}{10}$$

Bayesian Approach to Machine Learning

- MAP predicted 100% chance that third coin would be a head.
  - But the correct value was only 90% (obtained by marginalizing over θ).

- "Compute correct probability by marginalizing over parameters" is called the Bayesian approach to machine learning.
  - MAP approach optimizes posterior over parameter values.
    - Searches for the single "best" parameter value according to posterior.
  - Bayesian approach marginalizes posterior over parameter values.
    - Considers all possible parameter values, but upweighting ones with high posterior.

- MAP and Bayes are similar if posterior is "concentrated" at one θ.
  - But if there are many reasonable θ, Bayes can be much better.

Digression: Review of Independence

- Let A and B be random variables taking values $a \in \mathcal{A}$ and $b \in \mathcal{B}$.

- We say that A and B are independent if for all a and b we have:

$$p(a,b) = p(a)p(b)$$

- To denote independence of A and B we often use the notation:

$$A \perp\!\!\!\perp B$$

- The product of Bernoullis model assumes mutual independence:

$$x_i \perp\!\!\!\perp x_j \quad \text{for } \underbrace{\text{all } i \text{ and }}_{\text{this is the "mutual" part}} j$$

Digression: Review of Independence

- For independent A and B we have:

$$p(a \mid b) = \frac{p(a,b)}{p(b)} = \frac{p(a)p(b)}{p(b)} = p(a)$$

- We can also use this as a more intuitive definition:
  - A and B are independent if for all a and b where $p(b) \neq 0$ we have:

$$p(a \mid b) = p(a)$$

    - In words: "knowing b tells us nothing about a" (and vice versa: p(b | a)=p(b)).
    - This will often simplify calculations.

- Useful fact that can help determine if variables are independent:
  - $A \perp\!\!\!\perp B$ iff $p(a,b) = f(a)g(b)$ for some functions f and g.

Digression: Review of Conditional Independence

- We say that A is conditionally independent of B given C if:

$$p(a, b \mid c) = p(a \mid c)p(b \mid c) \quad \text{for all } a, b, \text{ and } c \text{ with } p(c) \neq 0$$

  – Same as independence definition, but "knowing extra stuff" C.

- Or, alternatively:

$$p(a \mid b, c) = p(a \mid c) \quad \text{or} \quad p(b \mid a, c) = p(b \mid c)$$

  – "If you know C, then *also* knowing B would tell you nothing about A."

- We often write this as: $A \perp\!\!\!\perp B \mid C$
- In naïve Bayes we assume $X_i \perp\!\!\!\perp X_j \mid Y$ for all i and j.
  – As we saw, this makes inference and learning easy.

Standard ML Independence Assumptions (MEMORIZE)

- In machine learning we typically make a standard set of independence assumptions:
  - IID assumption: training examples are independent of each other.

$$x^i \perp\!\!\!\perp x^j$$

    - "If you see example x^i, it doesn't tell you anything about x^j."
    - Maybe better framing is $x^i \perp\!\!\!\perp x^k \mid \mathcal{D}$: they're conditionally independent given the hidden "data-generating process" $\mathcal{D}$.
  - Independence of data given parameters.

$$x^i \perp\!\!\!\perp x^j \mid \Theta$$

    - "If we know the parameters, the examples are independent of each other"
    - Again, maybe better to think of this as $x^i \perp\!\!\!\perp x^k \mid \theta, \mathcal{D}$.
  - Independence of features X and parameters w in **discriminative** models.

$$w \perp\!\!\!\perp X$$

    - Discriminative models assume parameters are fixed, and w just transforms them to y (knowing X without y tells you nothing).
  - Conditional independence of data and hyper-parameters, given parameters:

$$X \perp\!\!\!\perp \alpha, \beta \mid \Theta$$

    - "Given the parameters, the hyper-parameters don't tell you anything more about the data."
- Later we'll discuss the models that lead to these assumptions, and testing independence in a model.

Bayesian Approach for Bernoulli-Beta Model

- Consider probability that $x^3 = 1$ after $x^1=1$ and $x^2=1$ with beta prior:

$$p(x^3=1 \mid X, \alpha, \beta) = \int_\theta p(x^3=1, \theta \mid X, \alpha, \beta)\, d\theta \quad \text{(marginalization rule)}$$

"posterior predictive"

$$= \int_\theta p(x^3=1 \mid \theta, X, \alpha, \beta)\, p(\theta \mid X, \alpha, \beta)\, d\theta \quad \text{(product rule)}$$

$$= \int_\theta \underbrace{p(x^3=1 \mid \theta)}_{\text{"prediction"}}\, \underbrace{p(\theta \mid X, \alpha, \beta)}_{\text{"posterior"}}\, d\theta \quad \text{(conditional independence)}$$

- Now use that posterior is a beta with parameters $\tilde{\alpha}$ and $\tilde{\beta}$.

$$= \int_\theta \theta\, p_\beta(\theta \mid \tilde{\alpha}, \tilde{\beta})\, d\theta \quad \text{(definition of Bernoulli and form of posterior)}$$

$$= \mathbb{E}[\theta] \quad \text{(expected value of } \theta \text{ under posterior distribution)}$$

$$= \frac{\tilde{\alpha}}{\tilde{\alpha} + \tilde{\beta}} \quad \text{(formula for expected value of } \theta \text{ under beta)}$$

Bayesian Approach for Bernoulli-Beta Model

- The correct probability of seeing a "head" after 2 flips in Bernoulli-beta:

$$p(x^3=1 \mid X, \alpha, \beta) = \int_0^1 p(x^3=1, \theta \mid X, \alpha, \beta)\, d\theta$$

$$= \frac{\tilde{\alpha}}{\tilde{\alpha}+\tilde{\beta}} \quad \text{(last slide)}$$

$$= \frac{n_1+\alpha}{(n_1+\alpha)+(n_0+\beta)}$$

- With a uniform prior, ($\alpha = \beta = 1$), then Pr(x^3 = 1 | x^1=1, x^2=1, α, β) = ¾.
  - The MAP under a uniform prior (which is MLE) would be θ =1.
    - It is less confident than MAP since it considers all possible θ values, not just the most likely.
    - Bayesian estimate is not degenerate even under a uniform prior here.
- Looks like Laplace smoothing, but trusts data less for same α and β.
  - For other models, MAP and Bayes can be much more different.

Effect of Prior in Bernoulli-Beta

- In Bayesian approach, hyper-parameters α and β can be thought of as "pseudo-counts".
  - The number of 0 and 1 outcomes you have in your imagination before you see any data.

- If we see 3 "heads" ($x^1=1,x^2=1,x^3=1$), the probability of a 4th under different priors:
  - Beta(1,1) prior is like seeing 1 imaginary head and 1 tail before flipping.
    - Probability is 4/5, even though all θ values under this uniform prior "equally likely".
  - Beta(3,3) prior is like seeing 3 imaginary heads and 3 tails.
    - Probability is 0.667. This is a stronger bias towards 0.5.
  - Beta(100,1) prior is like seeing 100 imaginary heads and 1 tail.
    - Probability is 0.990. This is a strong bias towards high θ values.
  - Beta(0.01,0.01) prior biases towards having an unfair coin (head or tail).
    - Probability is 0.997.

- We might hope to use an "uninformative" prior to not bias results.
  - We saw that with the "uniform" prior, Beta(1,1), it biases towards 0.5.
  - See bonus for additional details on why "uninformative" can be hard/ambiguous/impossible/undesirable.

Motivation: Controlling Complexity

- For many application, we need complicated models.
- But complex models can overfit.
- So what should we do?

- In CPSC 340 we saw two ways to reduce overfitting:
  - Model averaging (like in random forests).
  - Regularization (like in L2-regularized linear regression).

- Bayesian methods combine both of these.
  - Average over "models", weighted by posterior (which includes regularizer).
    - Recall that the regularizer corresponds to the negative logarithm of the prior.
  - This can allow you fit extremely complicated models without overfitting.

MAP vs Bayes for Categorical-Dirichlet

- MAP (regularized optimization) approach maximizes over parameters:

$$\hat{\Theta} \in \underset{\Theta}{\operatorname{argmax}}\{p(\Theta \mid X)\}$$
$$\equiv \underset{\Theta}{\operatorname{argmax}}\{p(X \mid \Theta)p(\Theta)\} \quad \text{(Bayes' rule)}$$
$$p(x=c \mid \hat{\Theta}) = \hat{\theta}_c$$

(I'm not explicitly including the conditioning on the hyper-parameters α)

- Bayesian approach predicts by integrating over possible parameters:

$$p(x=c \mid X) = \int_{\theta_1}\int_{\theta_2}\cdots\int_{\theta_K} p(x=c, \Theta \mid X)\, d\theta_K\, d\theta_{K-1}\cdots d\theta_1 \quad \text{(marg. rule)}$$
$$= \int_{\theta_1}\int_{\theta_2}\cdots\int_{\theta_K} p(x=c \mid \Theta, X)\, p(\Theta \mid X)\, d\theta_K\, d\theta_{K-1}\cdots d\theta_1 \quad \text{(product rule)}$$
$$= \int_{\theta_1}\int_{\theta_2}\cdots\int_{\theta_K} \theta_c \; p(\Theta \mid X)\, d\theta_K\, d\theta_{K-1}\cdots d\theta_1 \quad \text{(independence of data given parameters)}$$
$$\rightarrow \mathbb{E}[\theta_c] \quad \text{(mean of Dirichlet posterior)}$$

- Considers all possible Θ, and weights prediction by posterior for Θ.
  - Posterior contains a regularizer, so this is averaging and regularizing.

Ingredients of Bayesian Inference (MEMORIZE)

1. Likelihood $p(X \mid \Theta)$
   - Probability of seeing data given parameters.
2. Prior $\mathrm{p}(\Theta \mid \mathrm{A})$.
   - Belief that parameters are correct before we have seen data.
3. Posterior $\mathrm{p}(\Theta \mid \boldsymbol{X}, \mathrm{A})$.
   - Probability that parameters are correct after we have seen data.
   - MAP maximizes, but Bayesian approach uses the whole distribution.
4. Posterior predictive $p\left(\tilde{X} \mid \boldsymbol{X}, \mathrm{A}\right)$ (NEW).
   - Probability of new data $\tilde{X}$ given old data $\boldsymbol{X}$, integrating over parameters.
     - Specifically, we average the likelihood of $\tilde{X}$, weighted by the posterior of θ given $\boldsymbol{X}$.
   - Bayesian approach uses this distribution for inference.

Bayesian Approach: Discussion

- Our previous “learn then predict” approaches (MLE and MAP):
  - Optimize parameters θ (learning).
  - Do inference with the parameter estimate $\hat{\theta}$ (inference).

- Bayesian approach doesn’t really have a separate “learning phase”.
  - There is no optimization of the parameter θ.
  - You just skip to doing inference with the posterior predictive.
    - Consider all parameters θ.

- In practice, it often still looks like “learn then predict”.
  - Characterize the form of the posterior (“learning”).
  - Make predictions by doing integrals with the posterior (inference).

Bayesian Approach: Discussion

- The Bayesian approach is the optimal way to use a probabilistic model.
  - It's what the rules of probability say we should do.
  - …if you believe in your probability model (prior + likelihood).

- If the prior is bad, Bayesian approach can be harmful.
  - Bayesian approach historically criticized since it requires "subjective" prior.
  - But all models are based on "subjective" assumptions, sometime hidden!

- As we see more data, Bayesian posterior concentrates on MLE.
  - MLE/MAP/Bayes usually more or less agree for large datasets.

- Real problem with the Bayesian approach is that integrals are hard.
  - Posterior and posterior predictive only have a nice form with conjugate priors.
    - Otherwise, you need to use methods like Monte Carlo or "variational" methods for inference.

bonus!

Uninformative Priors and Jeffreys Priors

- We might want to use an uninformative prior to not bias results.
  - But this is often hard/impossible to do.
- We might think the uniform distribution, Beta(1,1), is uninformative.
  - But posterior will be biased towards 0.5 compared to MLE.
  - And if you use a different parameterization it won't stay uniform.
- We might think to use "pseudo-count" of 0, Beta(0,0), as uninformative.
  - But posterior isn't a probability until we see at least one head and one tail.
- Some argue that the "correct" uninformative prior is Beta(0.5,0.5).
  - This prior is invariant to the parameterization, which is called a Jeffreys prior.

Summary

- MLE for categorical distribution:
  - Write using unnormalized parameters and normalizing constant 'Z'.
- Dirichlet distribution:
  - "Probability distribution over discrete probability distributions".
  - When used as prior for categorical, posterior is also Dirichlet.
  - MAP estimate with Dirichlet prior gives generalization of Laplace smoothing.
- Conjugate prior:
  - Prior for a particular likelihood such that posterior is in same "family".
- Conditional independence of A and B [given C].
  - "Knowing A tells you nothing about B [if you also know C]".
  - Independence assumptions often simplify computations.
  - In ML we make a standard set of independence assumptions.
    - Data and hyper-parameters are independent given parameters.
- Bayesian learning.
  - Do inference with the posterior predictive (no "learning" phase).
  - Can be viewed as regularizing and averaging over parameters (harder to overfit).
  - Involves solving unpleasant integrals (unless you have a conjugate prior).

- Next time: priors on priors + relaxing IID.