# CPSC 440: Machine Learning

Bayesian Learning

Winter 2022

# Last Time: Bayesian Learning

- We contrasted the MAP vs Bayesian learning to making predictions:
  - For binary variable variables the two approaches can be written:

MAP:

Find $\hat{\theta} \in \text{argmax}\{\underbrace{p(\theta \mid X)}_{\text{"posterior"}}\}$

(compute $p(x=1 \mid \hat{\theta})$)

Bayesian:

$\underbrace{p(x=1 \mid X)}_{\text{"posterior predictive"}} = \int p(x=1, \theta \mid X)\, d\theta$

$= \int p(x=1 \mid \theta)\, \underbrace{p(\theta \mid X)}_{\text{"posterior"}}\, d\theta$

  - MAP makes predictions based only on the $\hat{\theta}$ with highest posterior.
  - Bayesian method weights all possible $\theta$ by their posterior.
- We discussed conjugate priors for a given likelihood.
  - Prior and posterior come from same "family" of distributions.
  - Often makes inference easier.

# 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:

$$A \perp B$$

- The product of Bernoullis model assumes mutual independence:

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

# 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:
  - $A \perp B$ 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:

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

  – Same as independence definition, but "knowing extra stuff" $C$.
- We can alternately use the more-intuitive definitions:

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

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

- We often write this as: $A \perp B \mid C$
- In naïve Bayes we assume $x_i \perp x_j \mid y$ for all 'i' and 'j'.
  – Which we saw 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.

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

    - "If you see example $x^i$, it does not make seeing example $x^j$ more likely."
      - I like to think of this as a conditional independence assumption, $x^i \perp x^k | \mathcal{D}$ (they are independent conditioned on the hidden "data-generating process" $\mathcal{D}$).
  - Independence of data given parameters.

    $$x^i \perp x^j \mid \theta$$

    - "If we know the parameters, the examples are independent of each other"
      - Again, I find this more intuitive if you think of this as $x^i \perp x^k | \theta, \mathcal{D}$.
  - Independence of features 'X' and parameters 'w' in discriminative models.

    $$w \perp X$$

    - 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:

    $$X \perp \alpha, \beta \mid \theta$$

    - "Given the parameters, the hyper-parameters do not tell you anything more about the data.
- Later we will 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:

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

"posterior predictive"

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

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

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

$$= \int_\theta \theta \; \mathrm{Beta}(\tilde{\alpha}, \tilde{\beta}) \qquad \text{(definition of Bernoulli and form of posterior)}$$

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

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

# 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 p($x^3$ = 1 | $x^1$=1, $x^2$=1, $\alpha$, $\beta$) = ¾.
  - The MAP under a uniform prior (which is MLE) would be $\theta$ =1.
    - It is less confident than MAP since it considers all possible $\theta$ values, not just the most likely.
- Looks like Laplace smoothing, but trusts data less for same $\alpha$ and $\beta$.
  - For other models, the difference between MAP and Bayes can be larger.

# Effect of Prior in Bernoulli-Beta

- In Bayesian approach, hyper-parameters $\alpha$ and $\beta$ 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 $\theta$ 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 $\theta$ values.
  - Beta(0.01,0.01) prior biases towards having an unfair coin (head or tail).
    - Probability is 0.997.
    - Called "improper" prior (does not integrate to 1), but posterior can be "proper".

- 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 see 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 allows you fit extremely-complicated models without overfitting.

# MAP vs Bayes for Categorical-Dirichlet

- MAP (regularized optimization) approach maximizes over parameters:

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

(I'm not explicitly including the conditioning on the hyper-parameters $\alpha$)

- Bayesian approach predicts by integrating over possible parameters:

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

- Considers all possible Θ, and weights prediction by posterior for Θ.
  - Posterior contains 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 $\mathrm{p}(\Theta \mid X, \mathrm{A})$.
   - Probability that parameters are correct after we have seen data.
   - MAP maximizes, but Bayesian approach uses the whole distribution.
4. Posterior predictive $p\left(\tilde{X} \mid X, \mathrm{A}\right)$ (NEW).
   - Probability of new data $\tilde{X}$ given old data $X$, integrating over parameters.
     - Specifically, we integrate the likelihood of $\tilde{X}$ times the posterior of $\theta$ given $X$.
   - Bayesian approach uses this distribution for inference.

# Bayesian Approach: Discussion

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

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

- 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 the prior.
  - It is what the rules of probability say we should do.
- Though if the prior is mis-leading, 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 agree as the data size increase.
- 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.

# Monte Carlo for Bayesian Inference

- Bayesian inference tasks usually involve integral parameters.
  - Where we compute some function 'g' times the posterior.

$$\int_{\theta} g(\theta)\, p(\theta \mid X, \alpha, \beta)\, d\theta = \mathbb{E}_{\underbrace{\theta \mid X, \alpha, \beta}}[g(\theta)]$$

Subscript shows distribution of random variable

  - For example, if $g(\theta) = p(\tilde{x} \mid \theta)$ we get the posterior predictive.
- If you can sample from the posterior, you can use Monte Carlo:
  1. Generate samples $\theta^1, \theta^2, \ldots, \theta^t$.
  2. Approximate the integral by: $\frac{1}{t} \sum_{i=1}^{t} g(\theta^i)$
- Sampling from the posterior is easy with standard conjugate priors.
  - We will discuss how to sample from continuous distributions later.

# Summary

- 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 (harder to overfit).
  - Involves solving unpleasant integrals (unless you have a conjugate prior).

- Next time: putting a prior on the prior and relaxing IID.

# 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.