CPSC 540: Machine Learning
Bayesian Statistics

Mark Schmidt

University of British Columbia

Winter 2019
For many of these tasks, we need very complicated models.

- We require multiple forms of regularization to prevent overfitting.

In 340 we saw two ways to reduce complexity of a model:

- Model averaging (ensemble methods).
- Regularization (linear models).

Bayesian methods combine both of these.

- Average over models, weighted by posterior (which includes regularizer).
Current Hot Topics in Machine Learning

Bayesian learning includes:
- Gaussian processes.
- Approximate inference.
- Bayesian nonparametrics.
Why Bayesian Learning?

- Standard L2-regularized logistic regression setup:
  - Given finite dataset containing IID samples.
  - E.g., samples \((x^i, y^i)\) with \(x^i \in \mathbb{R}^d\) and \(y^i \in \{-1, 1\}\).
  - Find “best” \(w\) by minimizing NLL with a regularizer to “prevent overfitting”.

\[
\hat{w} \in \arg\min_w \sum_{i=1}^{n} \log p(y^i | x^i, w) + \frac{\lambda}{2} \|w\|^2.
\]

- Predict labels of new example \(\tilde{x}\) using single weights \(\hat{w}\),
  \[\hat{y} = \text{sgn}(\hat{w}^T \tilde{x}).\]

  - But data was random, so weight \(\hat{w}\) is a random variables.
  - This might put our trust in a \(\hat{w}\) where posterior \(p(\hat{w} | X, y)\) is tiny.

- Bayesian approach: treat \(w\) as random and predict based on rules of probability.
Problems with MAP Estimation

Does MAP make the right decision?

Consider three hypotheses $\mathcal{H} = \{\text{"lands"}, \text{"crashes"}, \text{"explodes"}\}$ with posteriors:

\[
p(\text{"lands"} \mid D) = 0.4, \quad p(\text{"crashes"} \mid D) = 0.3, \quad p(\text{"explodes"} \mid D) = 0.3.
\]

The MAP estimate is "plane lands", with posterior probability 0.4.

- But probability of dying is 0.6.
- If we want to live, MAP estimate doesn’t give us what we should do.

Bayesian approach considers all models: says don’t take plane.

Bayesian decision theory: accounts for costs of different errors.
MAP vs. Bayes

- **MAP (regularized optimization) approach** maximizes over \( w \):
  \[
  \hat{w} \in \arg\max_w p(w \mid X, y)
  \equiv \arg\max_w p(y \mid X, w)p(w) \quad \text{(Bayes’ rule, } w \perp X)\]

- **Bayes’ approach** predicts by integrating over possible \( w \):
  \[
p(\tilde{y} \mid \tilde{x}, X, y) = \int_w p(\tilde{y}, w \mid \tilde{x}, X, y)dw \quad \text{marginalization rule}
  \]
  \[
  = \int_w p(\tilde{y} \mid w, \tilde{x}, X, y)p(w \mid \tilde{x}, X, y)dw \quad \text{product rule}
  \]
  \[
  = \int_w p(\tilde{y} \mid w, \tilde{x})p(w \mid X, y)dw \quad \tilde{y} \perp X, y \mid \tilde{x}, w
  \]

- **Considers all possible \( w \), and weights prediction by posterior for \( w \).**
Motivation for Bayesian Learning

Motivation for studying Bayesian learning:

1. **Optimal decisions** using rules of probability (and possibly error costs).
2. Gives estimates of **variability/confidence**.
   - E.g., this gene has a 70% chance of being relevant.
3. **Elegant approaches for model selection and model averaging**.
   - E.g., optimize $\lambda$ or optimize grouping of $w$ elements.
4. Easy to **relax IID assumption**.
   - E.g., hierarchical Bayesian models for data from different sources.
5. **Bayesian optimization**: fastest rates for some non-convex problems.
6. Allows models with **unknown/infinite number of parameters**.
   - E.g., number of clusters or number of states in hidden Markov model.

Why isn’t everyone using this?

- Philosophical: Some people don’t like “subjective” prior.
- Computational: Typically leads to nasty integration problems.
Bayesian statistics:
- Condition on the data, integrate (rather than maximize) over posterior.
- “All parameters are nuisance parameters”.

Next time: learning the prior?