Midterm Review

CMPUT 267: Basics of Machine Learning

Textbook Ch.1 - 7

Announcements/Comments

- A few updates are being made to the assignment to make it clearer, to be released tonight.
- If you have already started, do not worry! It does not change the assignment in any way, it just adds clarity.
- How was the practice midterm? It is longer than the quiz because (a) you are more used to this now and (b) you do not have to type.

Midterm Details

- The content is from Chapters 1 - 7
  - Chapter 7 is Introduction to Prediction problems
  - Chapter 8 is Linear Regression. Exam does not cover linear regression
- The exam only covers what is in the notes
- The focus is Chapters 4-7, but Chapter 1-3 are important background

Very brief summary of Ch 1-3

- Probability
- Estimators

Probability

- Define a **random variable**
- Define **joint** and **conditional probabilities** for continuous and discrete random variables
- Define **probability mass functions** and **probability density functions**
- Define **independence** and conditional independence
- Define **expectations** for continuous and discrete random variables
- Define **variance** for continuous and discrete random variables

Probability (2)

- Represent a problem probabilistically
    - e.g., how likely was the outcome?
- Use a provided distribution
    - I will always remind you of the density expression for a given distribution
- Apply **Bayes' Rule** to manipulate probabilities

Estimators

- Define **estimator**
- Define **bias**
- **Demonstrate that an estimator is/is not biased**
- Derive an expression for the variance of an estimator
- Define **consistency**
- Demonstrate that an estimator is/is not consistent
- Justify when the use of a **biased estimator** is **preferable**

Poll Question: When is the use of a biased estimator preferable?

- 1. It is always better because it biases towards the true solution
- 2. If the bias reduces the mean-squared error by reducing the variance
- 3. If the bias reduces the mean-squared error by reducing the variance
- 4. It is rarely justifiable

Answer: 2

Estimators (2)

- Apply concentration inequalities to derive **confidence bounds**
- Define **sample complexity**
- Apply concentration inequalities to derive sample complexity bounds
- Explain when a given concentration inequality can/cannot be used

Optimization

- Represent a problem as an optimization problem
- Solve a discrete problem by iterating over options and picking the one with the minimum value according to the objective
- Solve a continuous optimization problem by finding **stationary points**
    - **Poll: What is a stationary point?**

Poll Question: The following are true about stationary points

- 1. A stationary point is the global minimum of a function
- 2. A stationary point is a point where the gradient is zero
- 3. A global minimum is a stationary point, but a stationary point may not be a global minimum
- 4. If we find a stationary point, then we have found the minimum of our function
- 5. We can use the second derivative test to identify the type of stationary point we have

Answer: 2, 3 and 5

Optimization

- Represent a problem as an optimization problem
- Solve an analytic optimization problem by finding **stationary points**
- **Define first-order gradient descent**
- **Define second-order gradient descent**
- Define **step size** and **adaptive step size**
- Explain the role and importance of step sizes in first-order gradient descent
- Apply gradient descent to numerically find local optima

Exercise

- Imagine $c(w) = \frac{1}{2}(xw - y)^2$.
- What is the first-order update, assuming we are currently at point w_t?
  - i.e., the gradient descent update tells us how to modify our current point to descend on our surface c.

Answer: $w_{t+1} \leftarrow w_t - \eta_t c'(w_t)$ for some stepsize $\eta_t > 0$

$c'(w) = (xw - y)x$ so we have that. $w_{t+1} \leftarrow w_t - \eta_t(xw_t - y)x$

Exercise

- Imagine $c(w) = \frac{1}{2}(xw - y)^2$.
- What is the first-order update, assuming we are currently at point w_t?
  - i.e., the gradient descent update tells us how to modify our current point to descend on our surface c.
- What if instead we did $w_{t+1} \leftarrow w_t + \eta_t c'(w_t)$. What would happen?
- The second-order update is $w_{t+1} \leftarrow w_t - \dfrac{c'(w_t)}{c''(w_t)}$. Why might this update be preferable to the first-order? (poll)

Poll Question: Why might the second-order update be preferable?

- 1. It is easier to compute than the first-order one.
- 2. It tells us how to pick a good stepsize.
- 3. The second-order update is more likely to get stuck at a saddlepoint
- 4. The first-order update might get stuck in local minimum, but not the second-order update

Answer: 2

Parameter Estimation

- **Formalize a problem as a parameter estimation problem**
    - e.g., formalize modeling commute times as parameter estimation for a Poisson distribution, using maximum likelihood
- **Describe the differences between MAP, MLE, and Bayesian parameter estimation**
    - MAP $\max_\theta p(\theta \mid \mathcal{D})$ versus MLE $\max_\theta p(\mathcal{D} \mid \theta)$
    - Bayesian learns $p(\theta \mid \mathcal{D})$, reasons about plausible parameters
- Define a **conjugate prior**

The Likelihood Term and the Prior

- Likelihood:
  $p(\mathscr{D} \,|\, w) = \Pi_{i=1}^{n} p(x_i \,|\, w)$
- e.g., Poisson
  $$p(x_i \,|\, w) = \frac{w^{x_i} \exp(-w)}{x_i!}$$

- Prior:
  $p(w \,|\, \theta_0)$ for pdf or pmf parameters θ_0
- e.g., conjugate prior for Poisson is Gamma with parameters $\theta_0 = (a, b)$
  $$p(w \,|\, \theta_0) = \frac{w^{a-1} \exp(-w/b)}{b^a \Gamma(a)}$$

The Likelihood Term and the Prior

- Likelihood:
  $p(\mathscr{D} \mid w) = \Pi_{i=1}^{n} p(x_i \mid w)$
- e.g., Poisson
  $$p(x_i \mid w) = \frac{w^{x_i} \exp(-w)}{x_i!}$$

- MLE: maximize
  $p(\mathscr{D} \mid w) = \Pi_{i=1}^{n} p(x_i \mid w)$
- MAP: maximize
  $p(\mathscr{D} \mid w) p(w \mid \theta_0) = p(w \mid \theta_0) \Pi_{i=1}^{n} p(x_i \mid w)$

- Prior:
  $p(w \mid \theta_0)$ for pdf or pmf parameters θ_0
- e.g., conjugate prior for Poisson is Gamma with parameters $\theta_0 = (a, b)$
  $$p(w \mid \theta_0) = \frac{w^{a-1} \exp(-w/b)}{b^a \Gamma(a)}$$

The Likelihood Term and the Prior

- MLE: maximize
  $p(\mathscr{D} \mid w) = \Pi_{i=1}^{n} p(x_i \mid w)$
- MAP: maximize
  $p(\mathscr{D} \mid w) p(w \mid \theta_0) = p(w \mid \theta_0) \Pi_{i=1}^{n} p(x_i \mid w)$
- Bayesian: obtain posterior $p(w \mid \mathscr{D})$
- e.g., if Poisson likelihood with conjugate prior Gamma with prior parameters $\theta_0 = (a, b)$, then posterior is Gamma with $\theta_n = (a_n, b_n)$ where
  $a_n = a + \sum_{i=1}^{n} x_i$ and $b_n = \dfrac{1}{n + 1/b}$

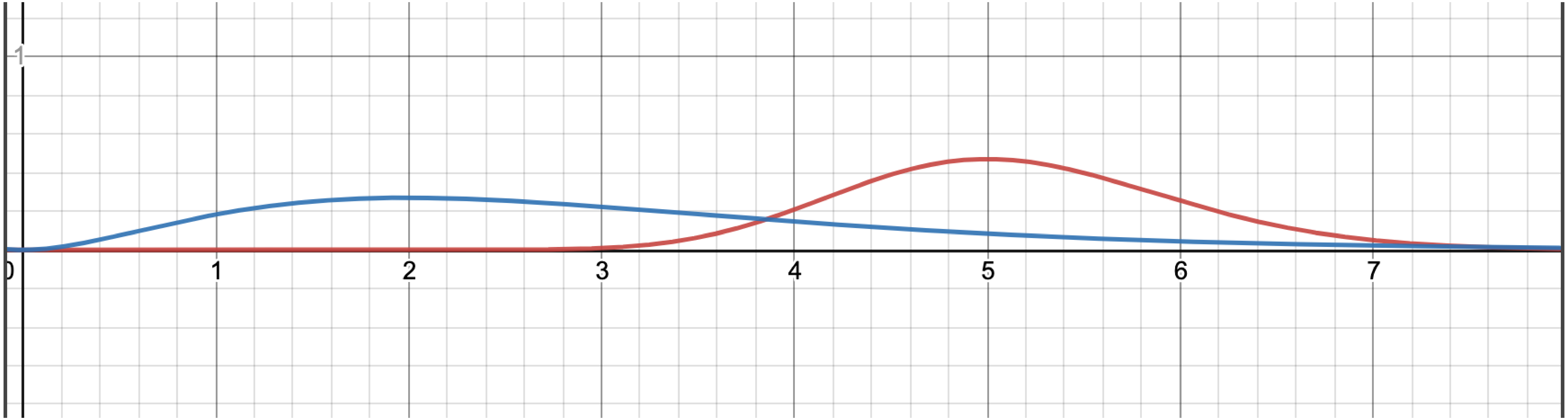
- Prior:
  $p(w \mid \theta_0)$ for pdf or pmf parameters θ_0
- e.g., conjugate prior for Poisson is Gamma with parameters $\theta_0 = (a, b)$
  $$p(w \mid \theta_0) = \frac{w^{a-1} \exp(-w/b)}{b^a \Gamma(a)}$$

Gamma Prior and Posterior

- For a = 3 and b = 1, we have $p(w) = \frac{1}{2} w^2 \exp(-w)$ because $\Gamma(3) = 2$
- For $\mathscr{D} = \{2,5,9,5,4,8\}$ we have $\sum_{i=1}^{n} x_i = 33$
- $a_n = a + \sum_{i=1}^{n} x_i = 36$ and $b_n = \frac{1}{n + 1/b} = 1/7$
- $p(w \mid \mathscr{D}) = \frac{w^{a_n - 1} \exp(-w/b_n)}{b_n^{a_n} \Gamma(a_n)} = \frac{w^{35} \exp(-7w)}{7^{-36} \Gamma(36)}$

Gamma Prior and Posterior

- For a = 3 and b = 1, we have $p(w) = \frac{1}{2} w^2 \exp(-w)$ as $\Gamma(k) = (k-1)!$
- $p(w \mid \mathscr{D}) = \dfrac{w^{a_n - 1} \exp(-w/b_n)}{b_n^{a_n} \Gamma(a_n)} = \dfrac{w^{35} \exp(-7w)}{7^{-36} \Gamma(36)}$ (Red)



Poll Question: Why is MAP useful, namely why is it useful to include a prior over the weights? (Select all that apply)

- 1. It incorporates bias to reduce the variance
- 2. The prior makes our solution closer to the true solution
- 3. It lets us reason about uncertainty in our parameters
- 4. It let's us incorporate expert knowledge about plausible weight values

Answer: 1, 4

Formalizing Prediction

- **Supervised learning problem:** Learn a **predictor** $f : \mathcal{X} \to \mathcal{Y}$ from a dataset $\mathcal{D} = \left\{(\mathbf{x}_i, y_i)\right\}_{i=1}^{n}$
  - $\mathcal{X}$ is the set of **observations**, and $\mathcal{Y}$ is the set of **targets**
- **Classification** problems have discrete, unordered targets
- **Regression** problems have continuous targets
- Predictor performance is measured by the **expected** $\text{cost}(\hat{y}, y)$ of predicting $\hat{y}$ when the true value is y
- An **optimal predictor** for a given distribution **minimizes** the expected cost

Prediction Concepts

- Describe the differences between **regression** and **classification**
- **Derive the optimal classification predictor for a given cost**
- Derive the **optimal regression predictor** for a given cost
- Understand that the optimal predictor is different depending on the cost
- Describe the difference between **irreducible** and **reducible error**
  - Even an optimal predictor has some **irreducible error**. **Suboptimal** predictors have additional, **reducible error**

$$\mathbb{E}[C] = \underbrace{\boxed{\mathbb{E}\left[\left(f(X) - f^*(X)\right)^2\right]}}_{\text{Reducible error}} + \underbrace{\boxed{\mathbb{E}\left[\left(f^*(X) - Y\right)^2\right]}}_{\text{Irreducible error}}$$

Is Cost the Same as our Objective c?

- We gave this a **different name** to indicate it might not be
- The **Cost** is the penalty we incur for inaccuracy in our predictions
- We parameterize our function or distribution with parameters $\mathbf{w}$
- Our **objective** to find $\mathbf{w}$ has typically been the negative log likelihood
- Example: we might learn $p(y \mid \mathbf{x}, \mathbf{w})$ using $c(\mathbf{w}) = -\ln p(\mathscr{D} \mid \mathbf{w})$
- For the **0-1 cost,** we **evaluate** the predictor $f(\mathbf{x}) = \arg\max_y p(y \mid \mathbf{x}, \mathbf{w})$

Any Questions?

- Switch now to going over the practice midterm