Midterm Review

CMPUT 267: Basics of Machine Learning

Textbook Ch.1 - 7

Announcements

- Bayesian Linear Regression chapter updated, to add some clarifying details
- Hybrid Delivery for Midterm: you can choose to write it online (like the quiz) or in person (in the classroom)
- Please fill out the Discord poll again asking for preferences
- **The final exam will be in-person. Please email me asap if you cannot do it in person (e.g., due to visa issues)**

Midterm Rules

- See the Exam Instructions link
- If online, you **must** join Zoom and have your camera on
- The exam is open book but you cannot use the internet
    - Everything must be downloaded ahead of time
    - In class you can download materials onto a device, but cannot be connected to the internet
- **New explicit rule: You cannot use any translation services**

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

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**

Go to menti.com and use code 3836 4159

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?**

Go to menti.com and use code 3836 4159 or https://www.menti.com/bzhs3fj22o

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?

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**

Prediction

- 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**

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

Reducible error

Irreducible error

Summary slide for 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 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
- Even an optimal predictor has some **irreducible error**.
  Suboptimal predictors have additional, **reducible 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 $c(\mathbf{w}) = -\ln p(\mathcal{D} \mid \mathbf{w})$
- For the **0-1 cost,** we **evaluate** the predictor $f(\mathbf{x}) = \arg\max_y p(y \mid \mathbf{x}, \mathbf{w})$
- For the medical costs example, we derived a different predictor f in class

Any Questions?

- Switch now to going over the practice midterm(s)