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 increasing 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**
  - A point w is a stationary point if $c'(w) = 0$
  - or for multivariate $\mathbf{w}$, $\nabla c(\mathbf{w}) = 0$

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 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

Closed-form solutions

- $c(w) = (w - 3)^2$ has a closed-form solution because

$$c'(w) = 2(w-3) = 0 \implies w - 3 = 0 \implies w = 3.$$

- $c(w) = w^2 + \exp(w)$ does not have a closed-form solution because

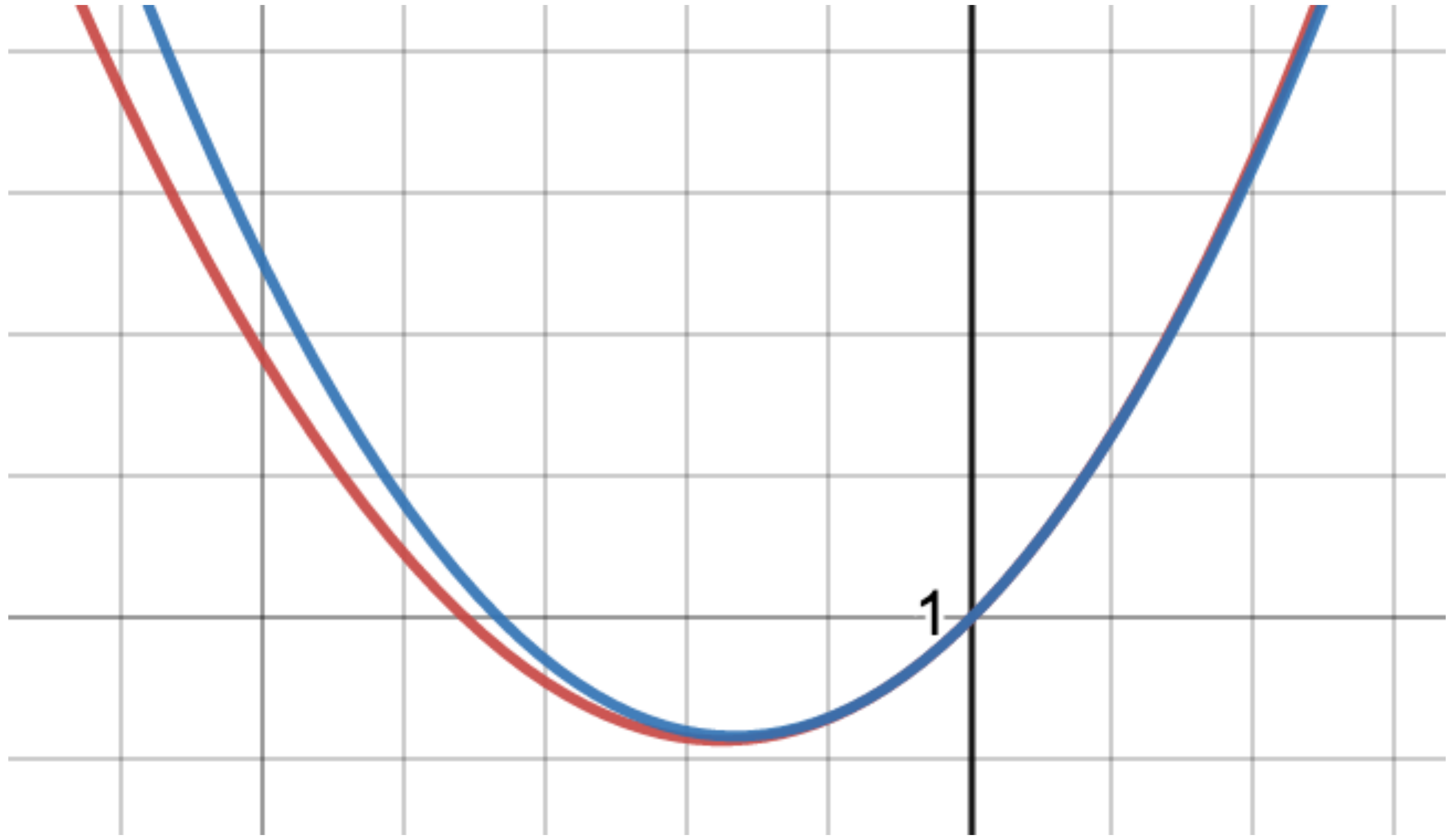
$$c'(w) = 2w + \exp(w) = 0 \implies \exp(w) = -2w$$

- Can't isolate w on one side, to get an explicit formula (closed-form)
  - Note: this c is not a hard optimization problem, it is convex

Second-order update

Example 14: Let us revisit our example $c(w) = w^2 + \exp(w)$, where $c'(w) = 2w + \exp(w)$ and $c''(w) = 2 + \exp(w)$. Let us start $w_0 = 0$ and do one second-order update.

$$
\begin{aligned}
w_1 &= w_0 - \frac{c'(w_0)}{c''(w_0)} \\
&= 0 - \frac{0 + \exp(0)}{2 + \exp(0)} \\
&= -\frac{1}{3}
\end{aligned}
$$



red line is c(w),
blue line is second-order Taylor approximation
around w = 0

Now let us do the next update.

$$
\begin{aligned}
w_2 &= w_1 - \frac{c'(w_1)}{c''(w_1)} \\
&= -\frac{1}{3} - \frac{-\frac{2}{3} + \exp(-\frac{1}{3})}{2 + \exp(-\frac{1}{3})} \\
&= -0.3516893316
\end{aligned}
$$

$$
\begin{aligned}
\hat{c}(w) &= c(w_0) + (w - w_0)c'(w_0) + \frac{1}{2}(w - w_0)^2 c''(w_0) \\
&= \exp(0) + w\exp(0) + (2 + \exp(0))\frac{1}{2}w^2 = 1 + w + \tfrac{3}{2}w^2
\end{aligned}
$$

Stochastic gradient descent

- If $c(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} c_i(\mathbf{w})$, then we can be more computationally efficient by using a stochastic approximation to the gradient on each step
- Each update consists of taking a mini-batch $\mathscr{B}$ and updating with
- $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta_t \frac{1}{b} \sum_{i \in \mathscr{B}} \nabla c_i(\mathbf{w}_t)$

Stochastic gradient descent

- Each update consists of taking a mini-batch $\mathscr{B}$ and updating with
- $$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta_t \frac{1}{b} \sum_{i \in \mathscr{B}} \nabla c_i(\mathbf{w}_t)$$
- We do this for T iterations (where T is likely more than the number of iterations used for GD)
- Example, if T = 640, n = 4096 and the mini-batch size is b = 32, then we need to do numepochs = 5 to get T = (n/b)*numepochs = 640 updates

You do not need to know

- Specific step-size selection algorithms
  - Adagrad
  - Line search
- I won’t get you to tell me about stopping criteria, for GD or SGD
  - for GD we usually check if the gradient norm becomes small enough
  - for SGD we just fixed the number of epochs (in practice, you might periodically check if improvement in the objective function has plateaued)

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 $\max_w p(w \mid \mathcal{D})$ versus MLE $\max_w p(\mathcal{D} \mid w)$
  - Bayesian learns $p(w \mid \mathcal{D})$, reasons about plausible parameters
- Define a **conjugate prior**

The Likelihood Term and the Prior

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

- 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

- 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 $\theta_0 = (a, b)$

$$p(w \mid \theta_0) = \frac{w^{a-1} \exp(-w/b)}{b^a \Gamma(a)}$$

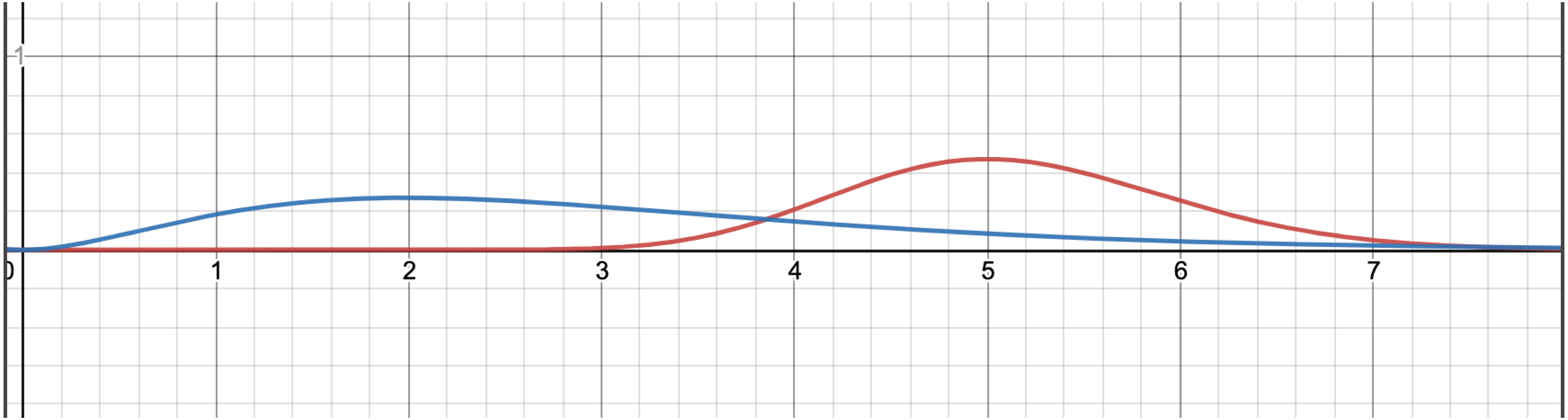
Gamma Prior and Posterior

$$p(w \mid \theta_0) = \frac{w^{a-1}\exp(-w/b)}{b^a\Gamma(a)}$$

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

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)



What is not a conjugate prior?

- Assume p(x) is Poisson.
- Imagine we pick the prior p(w) to be a uniform distribution on [1, 5], reflecting that we are 100% sure the average number of accidents is between 1 and 5 for the factory (before seeing data)
  - but we have no idea what the average is beyond that, all equally likely
- Then the posterior is just some integral we cannot solve

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

You do not need to know

- Any specific conjugate priors, or specific formulas for pmfs/pdfs
  - I will tell you if something is a conjugate prior, you just need to know what that means
- I will not get you to do complex derivations, to solve MLE or MAP

Formalizing Prediction

- **Supervised learning problem:** Learn a **predictor** $f : \mathscr{X} \rightarrow \mathscr{Y}$ from a dataset $\mathscr{D} = \left\{(\mathbf{x}_i, y_i)\right\}_{i=1}^{n}$
    - $\mathscr{X}$ is the set of **observations**, and $\mathscr{Y}$ 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

Difference between Classification and Regression

- If I learn a classifier f(x), for classes {0, 1, 2, 3}, what is the range of the predictor f?
- What is the optimal predictor for 0-1 cost for classification?
- Can I use classes like {apples, oranges, pineapples}? How would we write our optimal predictor for this set of classes?
- What is the optimal prediction for squared error costs for regression?

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**

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

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 $f(\mathbf{x}) = \arg\max_y p(y \mid \mathbf{x}, \mathbf{w})$

Optimal predictors vs MLE/MAP

- Why do we learn $p(y \mid \mathbf{x})$ if we only care about $\mathbb{E}[Y \mid x]$?
- Why do we have to learn a predictor $f(\mathbf{x})$ that returns one prediction $\hat{y}$ instead of just learning $p(y \mid \mathbf{x})$ and returning the whole distribution?
- Is the optimal predictor an MLE or MAP estimator?

Optimal predictors vs MLE/MAP

- Why do we learn $p(y \mid \mathbf{x})$ if we only care about $\mathbb{E}[Y \mid x]$?
  - We still want to recognize that y is stochastic for a given x, so we reason about $p(y \mid \mathbf{x})$ and about modelling it
  - For regression, we don't need $p(y \mid \mathbf{x})$, but we do for other predictors
- Why do we have to learn a predictor $f(\mathbf{x})$ that returns one prediction $\hat{y}$ instead of just learning $p(y \mid \mathbf{x})$ and returning the whole distribution?
  - At some point you have to make a decision: are you going to treat or not?
- Is the optimal predictor an MLE or MAP estimator?
  - The optimal predictor f* has nothing to do with data. We learn f on data (using MAP or MLE) to try to best approximate f*. Chapter 7 is not about learning nor data

Any Questions?

- Switch now to going over the practice midterm