Loss and Risk - Midterm 1 Review

Data 100, Fall 2019

Suraj Rampure

Monday, October 7th, 2019

Agenda

- Loss functions
- Risk vs. empirical risk
- Minimizing risk with and without calculus
- Practice problems

This will end up being a review of a good portion of Discussion 2 and 3.

This will be posted on Piazza, and additionally at
http://surajrampure.com/teaching/ds100.html

Loss Functions

Suppose we have a collection of data points $\{x_1, x_2, \ldots, x_n\}$, and we want to come up with a **summary statistic** c for this data, that is the "best", in some sense.

- Prediction error: $x_i - c$ actual−prediction
- To determine the "best" c, we need a function in terms of our true value x_i and prediction c, that increases as our error increases

L_2 (i.e. "squared") loss for a single point: $L_2(x_i, c) = (x_i - c)^2$

L_1 (i.e. "absolute") loss for a single point: $L_1(x_i, c) = |x_i - c|$

Risk vs. Empirical Risk

Risk is defined as the **expected loss** over *all possible datasets*, i.e.

$$\mathbb{E}[L(X, c)]$$

- Since we don't have access to *all possible datasets*, we represent our data as a random variable X.

Empirical risk, then, is the **average loss** over *the dataset we have*.

$$\frac{1}{n}\sum_{i=1}^{n} L(x_i, c)$$

- First, we will look at minimizing empirical risk. We'll then switch over to the random variable context and compare our results.

Minimizing Empirical Risk with Squared Loss

Let's consider the optimization problem

Empirical risk

$$\min_c \underbrace{\frac{1}{n} \sum_{i=1}^{n} (x_i - c)^2}$$

There are two approaches we can take:

$\hat{c}$: optimal value of c

1. Find the minimizing $\hat{c}$ using calculus
2. Using a few algebraic tricks

$$R(c) = \frac{1}{n} \sum_{i=1}^{n} (x_i - c)^2$$

$$\frac{dR(c)}{dc} = \frac{1}{n} \sum_{i=1}^{n} 2(x_i - c)(-1) = 0$$

$$-\frac{2}{n} \sum_{i=1}^{n} (x_i - c) = 0$$

$$\sum_{i=1}^{n} (x_i - c) = 0$$

$$\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} c = 0$$

$$\underbrace{c + c + \cdots + c}_{n \text{ times}}$$

$$\boxed{\frac{1}{n} \sum_{i=1}^{n} x_i = \hat{c}} = \bar{x}$$

Sample Mean Minimizes Empirical Risk (in this case)

In short: The sample mean minimizes empirical squared loss.

$$R(c) = \frac{1}{n} \sum_{i=1}^{n} (x_i - c)^2$$

$$\hat{c} = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

The **minimum value**, i.e. the empirical risk when $c = \hat{c}$, is the **sample variance**!

$$R(\hat{c}) = R(\bar{x}) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \text{sample var}$$

Minimizing Empirical Risk with Absolute Loss

Now, let's consider the optimization problem

$$\min_c \overbrace{\frac{1}{n}\sum_{i=1}^{n}|x_i - c|}^{R(c)}$$

$$|x| = \begin{cases} x & \text{if } x \geq 0 \\ -x & \text{if } x < 0 \end{cases}$$

$$|x_i - c| = \begin{cases} x_i - c & x_i \geq c \\ c - x_i & x_i < c \end{cases}$$

$$\frac{d|x_i - c|}{dc} = \begin{cases} -1 & x_i \geq c \\ 1 & x_i < c \end{cases}$$

$$\frac{dR(c)}{dc} = \frac{1}{n}\sum_{i=1}^{n}\frac{d|x_i - c|}{dc}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(1 + (-1) + (-1) + \dots\right)$$

$$0 = \frac{1}{n}\left[(1)(\#\, x_i < c) + (-1)(\#\, x_i \geq c)\right]$$

$$\boxed{(\#\, x_i < c) = (\#\, x_i \geq c)}$$

$$\Rightarrow \quad \hat{c} = \text{median}(\{x_1, \dots, x_n\})$$

<u>Extra</u> : Consider the approach from Discussion 3,

where m_c values are $\leq c$

(and thus $n-m_c$ values are $> c$)

$$\Rightarrow \quad (\#\, x_i \leq c) = (\#\, x_i > c)$$

$$m_c = n - m_c$$

$$2m_c = n$$

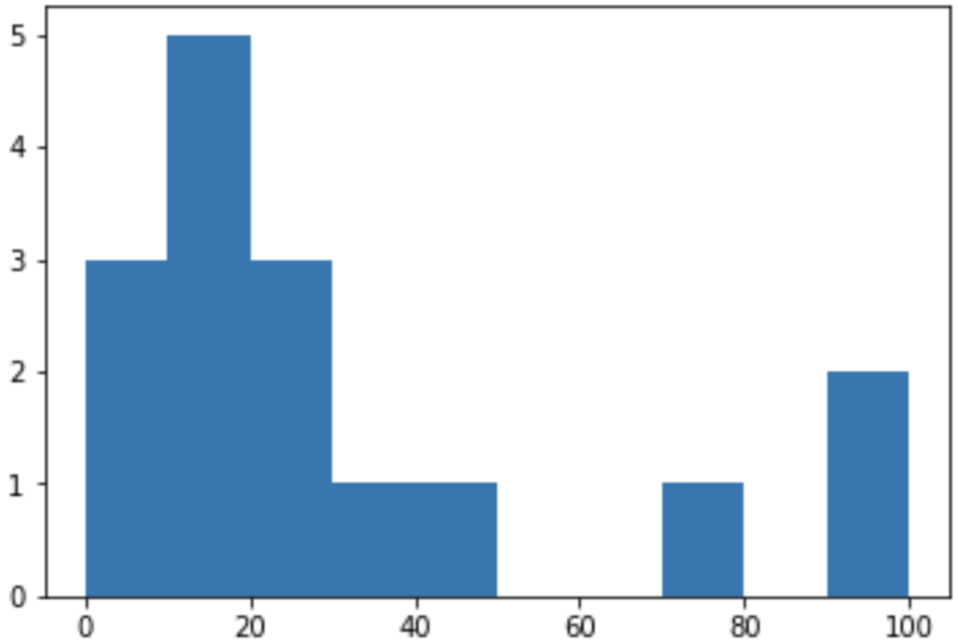
$$m_c = \frac{n}{2} \longrightarrow$$ i.e. half of values are above, half are below

L_1 vs. L_2

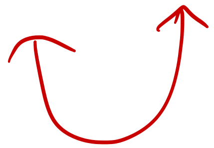
Consider the following set of points:

```
pts = np.array([0, 0, 5, 10, 10, 15, 15, 15, 20, 20, 25, 30, 40, 70, 90, 100])
print("mean: ", np.mean(pts))
print("median: ", np.median(pts))
plt.hist(pts);
```

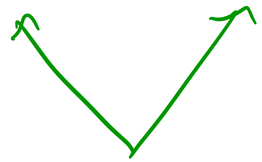
```
mean:  29.0625
median:  17.5
```



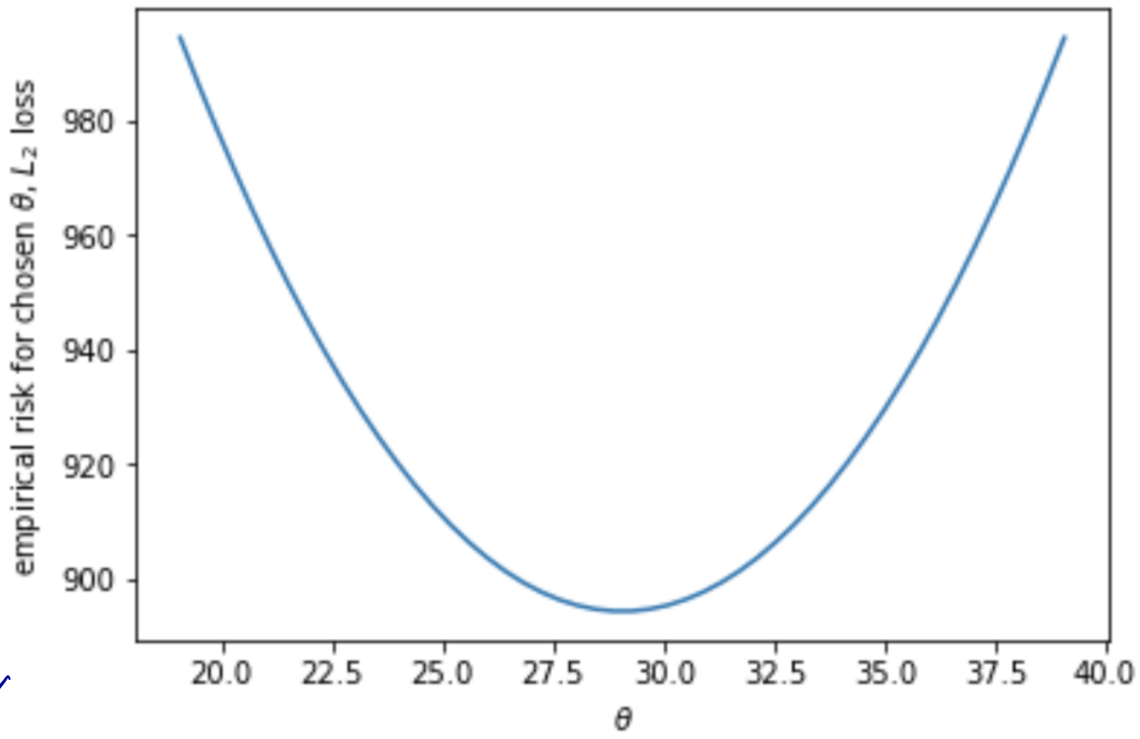
L_2 loss for a single pt

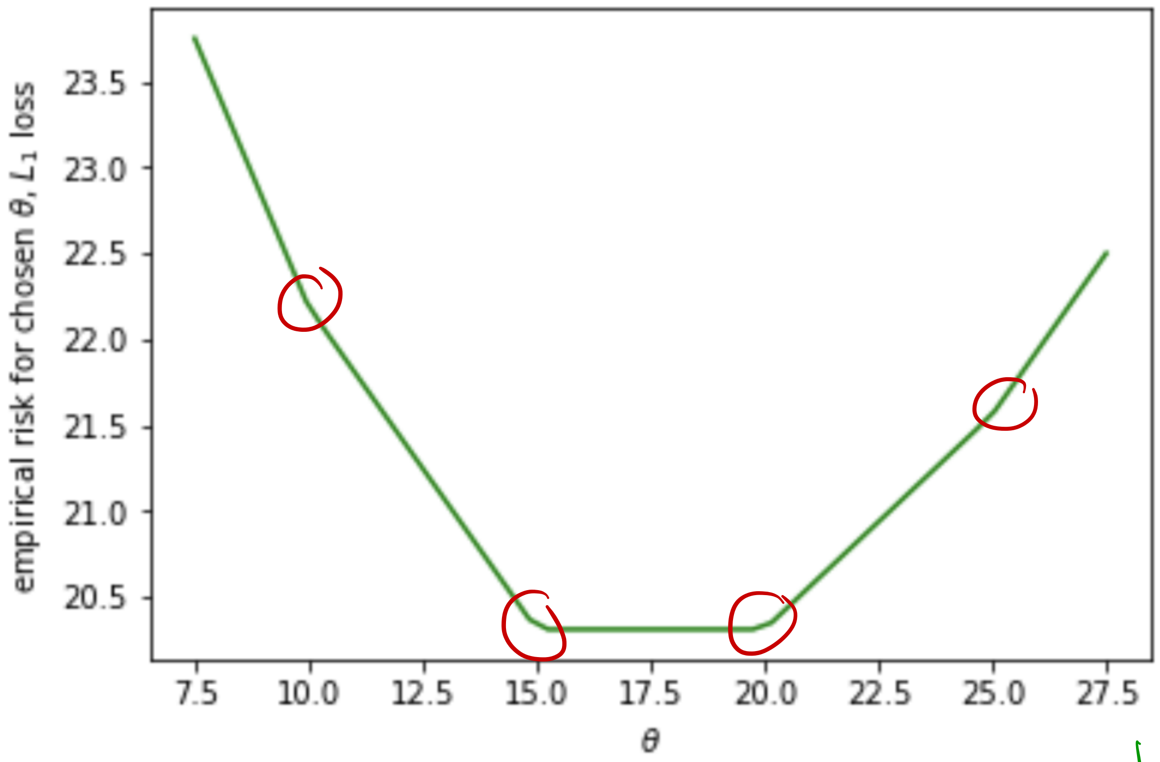


L_1 loss for a single pt



Let's look at plots of the empirical risk for both L_2 and L_1 loss, with varying values of θ, to see if our findings were correct.





Some questions to consider:

- Why are the optimal values of θ so different in the two cases? mean vs. median
- Why is the plot for squared loss smooth, but the plot for absolute loss so "choppy"?
- In what situations would we use squared loss? Absolute loss? Disc 3, #1

$$\frac{1}{n}\left((\theta-0)^2+(\theta-5)^2+(\theta-15)^2+\dots\right)$$

sum of quadratics is a quadratic

$$\frac{1}{n}\left[|\theta-0|+|\theta-5| + |\theta-15|+\dots\right]$$

sum of abs is not a single abs!

Sample Problem 1 (adapted from Fall 2018's midterm)

Let's define a custom loss function called the "OINK" loss:

$$L_{OINK}(x_i, c) = \begin{cases} a(c - x_i) & c \geq x_i \\ b(x_i - c) & c < x_i \end{cases}$$

Consider the set of values $\{0, 10, 20, 30, 40, 50, 60\}$. Determine the optimal $\hat{c}$ that minimizes empirical risk in each of the following cases:

$(\# \, x_i \leq c) = \frac{b}{a}(\# \, x > c)$

1. $a = b = 1$ $\hat{c}$ = median = 30
2. $a = 1, b = 5$ $\hat{c} = 50$

   $(\# \, x_i \leq c) = 5(\# \, x_i > c)$
3. $a = 3, b = 6$ $\hat{c} = 40$
4. For arbitrary a, b (this is more conceptual --- what exactly is happening?)

$\rightarrow \quad \hat{c} = 100 \cdot \frac{b}{a+b}$ percentile

Minimizing Risk with Squared Loss

Now, let's switch gears and consider **risk**, not just empirical risk. We can theoretically look at risk with any loss function, but we tend to consider L_2:

$$R(c) = \mathbb{E}[(X - c)^2]$$

Again, there are two approaches to finding this minimum value.

Minimizing Risk with Squared Loss, using Calculus

One approach is to use calculus:

$$R(c) = \mathbb{E}[(X - c)^2]$$

$$R(c) = \mathbb{E}[X^2 - 2cX + c^2]$$

$$R(c) = \mathbb{E}[X^2] - 2c\mathbb{E}[X] + c^2$$

$$\Rightarrow \frac{dR(c)}{dc} = -2\mathbb{E}[X] + 2c = 0$$

$$\Rightarrow \hat{c} = \mathbb{E}[X]$$

Minimizing risk with Squared Loss, without Calculus

Note: $X - c = (X-\mu) + (\mu - c)$, $\mu = E[X]$ $\qquad (\square + \Delta)^2 = \square^2 + 2\square\Delta + \Delta^2$

Then,

$$E[(X-c)^2] = E\Big[\big(\underbrace{(X-\mu)}_{\square} + \underbrace{(\mu-c)}_{\Delta}\big)^2\Big]$$

$$= E\left[(X-\mu)^2 + 2(X-\mu)(\mu-c) + (\mu-c)^2\right]$$

$$= \underbrace{E\left[(X-\mu)^2\right]}_{\text{definition of var}(X)} + E\left[\underbrace{2}_{\text{constant}}(X-\mu)\underbrace{(\mu-c)}_{\text{constant}}\right] + E\left[\underbrace{(\mu-c)^2}_{\text{constant}}\right]$$

$$= \text{var}(X) + 2(\mu-c)\underbrace{(E[X]-\mu)}_{=0} + (\mu-c)^2$$

$$= \text{var}(X) + (\mu-c)^2$$

→ var(X) independent of c

→ $(\mu-c)^2$ minimized at $c = \mu = E[X]$

Expectation Minimizes Risk (in this case)

- Previously, we saw that the **sample variance** was the minimum value (output) of empirical risk with squared loss, with the optimal value (input) being the **sample mean**.
- A similar property holds true when looking at risk.

$$R(c) = \mathbb{E}[(X - c)^2]$$

- The value that minimizes $R(c)$ is $\hat{c} = \mathbb{E}[X]$
- The minimum value of $R(c)$ is the variance of X, i.e.

$$\underbrace{\mathbb{E}[(X - \mathbb{E}[X])^2]}_{var(X)}$$

Sample Problem 2 (adapted from Spring 2018's final)

Suppose we observe a sample of n runners from a larger population, and we record their race times $X_1, X_2, ..., X_n$. We want to estimate the **maximum race time** θ^* in the population. When comparing estimates, we prefer whichever is closer to θ^* without going over. We consider the following three estimators based on our sample:

- $\hat{\theta}_1 = \max_i X_i$
- $\hat{\theta}_2 = \frac{1}{n}\sum_i X_i$
- $\hat{\theta}_3 = \max_i X_i + 1$

Essentially, want to get as close to max of population, without exceeding

a) True or False: $\hat{\theta}_1$ is never an overestimate, but could be an underestimate of θ^*. True : max(sample) $\leq$ max(pop)

b) True or False: $\hat{\theta}_1$ is never a worse estimate of θ^* than $\hat{\theta}_2$. True : $\hat{\theta}_1 \geq \hat{\theta}_2$

c) True or False: $\hat{\theta}_3$ is never a worse estimate of θ^* than $\hat{\theta}_1$. False : could be an overestimate

d) Which loss $l(\hat{\theta}, \theta^*)$ best reflects our goal of *"closest without going over"*?

$$l_A(\hat{\theta}, \theta^*) = (\hat{\theta} - \theta^*)^2$$

$$l_B(\hat{\theta}, \theta^*) = \begin{cases} \theta^* - \hat{\theta} & \hat{\theta} \leq \theta^* \\ \infty & \text{else} \end{cases}$$

$$l_C(\hat{\theta}, \theta^*) = |\hat{\theta} - \theta^*|$$

$$l_D(\hat{\theta}, \theta^*) = \begin{cases} \theta^* - \hat{\theta} & \hat{\theta} \leq \theta^* \\ 0 & \text{else} \end{cases}$$

$\hat{\theta}$

If our guess $\leq \theta^*$, penalize the difference

If our guess $> \theta^*$, very bad!

Sample Problem 3 (adapted from Fall 2017's practice final)

Suppose we observe a dataset $\{x_1, x_2, \ldots, x_n\}$ of independent and identically distributed samples from the exponential distribution.

$$L(\lambda) = -n \log(\lambda) + \lambda \sum_{i=1}^{n} x_i$$

Determine the parameter value λ that minimizes the above loss function.

Note: I've intentionally removed a lot of detail from this problem, as it's not quite presented the same way we'd present a problem in Fall 2019. This is primarily to serve as mechanical practice.

Treat this as a calculus problem: ignore the meaning for now.

$$L(\lambda) = -n\log(\lambda) + \lambda\sum_{i=1}^{n} x_i$$

$$\frac{dL(\lambda)}{d\lambda} = \frac{-n}{\lambda} + \sum_{i=1}^{n} x_i = 0$$

$$\frac{n}{\lambda} = \sum_{i=1}^{n} x_i$$

$$\Rightarrow \boxed{\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}}}$$

Good luck!