INDR 450/550

Spring 2022

Lecture 2: Intro cont., estimation, simple forecasts

Feb. 16, 2022

Fikri Karaesmen

Announcements

- Blackboard page is becoming active
  - Slides of first lecture
  - Links to boks
- We may start looking at some data today
  - Next week, I’ll upload a video on basic analysis of some forecasting models

A typical operational problem

- Machine learning (in the supervised learning framework) starts with data $(\mathbf{y_1}, \mathbf{x_1}), (\mathbf{y_2}, \mathbf{x_2}) \ldots (\mathbf{y_n}, \mathbf{x_n})$ and focuses on the prediction problem of $\mathbf{Y}|\mathbf{X}$
- and proposes a number of effective tools.
- On the other hand, prescriptive analytics focuses on:

$$\min_{\mathbf{z}} E[c(\mathbf{Y}|\mathbf{X} = \mathbf{x}, \mathbf{z})]$$

- and of course also on finding the minimizer $\mathbf{z}^*$.
- Note that the typical ML-based problem is also an optimization problem where some error function is minimized.
- Prescriptive analytics therefore considers such nested optimization problems one for estimation, the other on operational cost minimization.

Example:

- Let Y be a uniformly distributed random variable in $(0,1)$ and $c(y) = y^2$.

$$E[c(Y)] = E[Y^2] = \int_0^1 y^2 \, dy = \frac{1}{3}.$$

whereas:

$$c(E[Y]) = E[Y]^2 = \left(\frac{1}{2}\right)^2 = \frac{1}{4}.$$

Remark

- Recall that we started by defining a standard problem:

$$\min_{\mathbf{z}} E[c(\mathbf{Y}, \mathbf{z})]$$

  where $\mathbf{z}$ is a decision variable and $\mathbf{Y}$ is a random variable.
- Please note that

$$E[c(\mathbf{Y}, \mathbf{z})] \neq c(E[\mathbf{Y}], \mathbf{z})$$

- If the two were equal, then we could leverage ML methods to estimate $E[\mathbf{Y}|\mathbf{X}]$ and would solve a deterministic optimization problem.

The Newsvendor Problem

- A single-period random demand inventory problem (the newsvendor problem). We have to order a quantity in advance of the demand realization.
- No opportunity to reorder during the sales season, unsatisfied demand is lost
- Unsold items are salvaged at a value below their purchasing cost.
- Since demand is not known with certainty, there will be a mismatch between the supply and demand.
- Assume that we somehow know the distribution of random demand D. We can the maximize the expected profit:

$$\max_q E\left[-cq + p\min(q, D) + s(q-D)^+\right]$$

p: sales price, c: purchase cost, s: salvage value and p>c>s.

The Newsvendor Problem

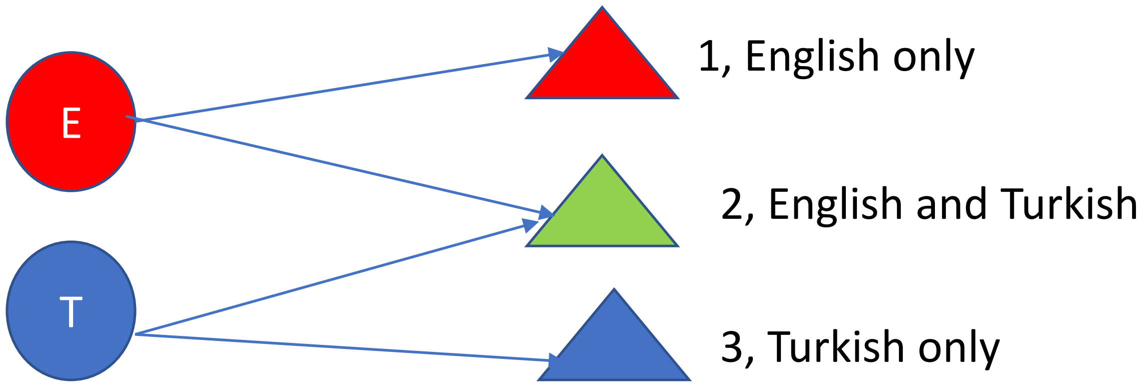
- In practice, we might have data that are past observations of realized demand $d_1, d_2, ..., d_n$.
- We then have two basic alternatives i) fit a probability distribution to the data and obtain the corresponding random variable D ii) Use the sample as our 'world' and perform empirical optimization.This is called sample average approximation (and empirical risk minimization in ML).
- We assign a weight that equals $1/n$ to each observation and solve the following deterministic problem

$$\max_q \; -cq + \frac{\sum_{i=1}^{n} p\min(q, d_i) + s(q - d_i)^+}{n}$$

- Note that the solution of the above problem finds the optimal order quantity that would maximize the average profit for the sample.

Flexible Capacity Design

- A more complicated version of the newsvendor problem: flexible capacity design.
- Assume that you are staffing a call center that responds to calls in Turkish and English.
- You have three types of agents: only Turkish speaking (low cost), only English speaking (medium cost) and speaking both Turkish and English (high cost).

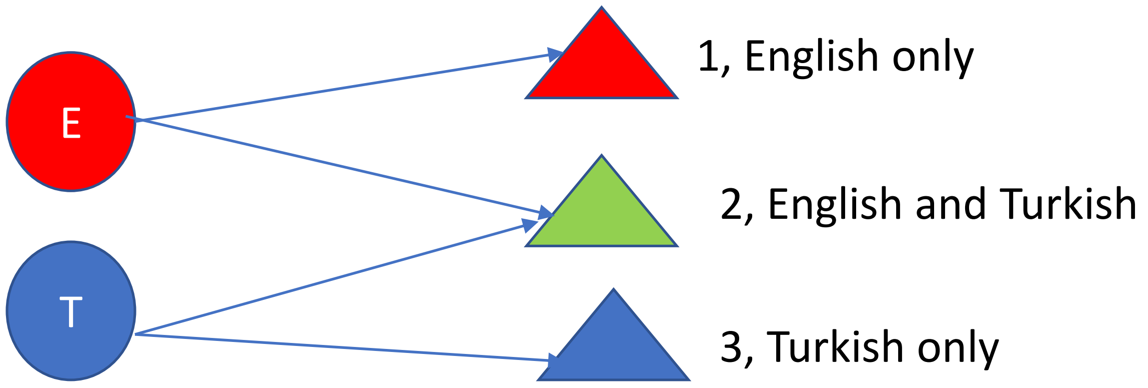


Flexible Capacity Design

- How many agents of each type to staff given your demand prediction?

$$\max_{q_1,q_2,q_3} -\sum_{j=1}^{3} c_j q_j + p(X_E + X_T)$$

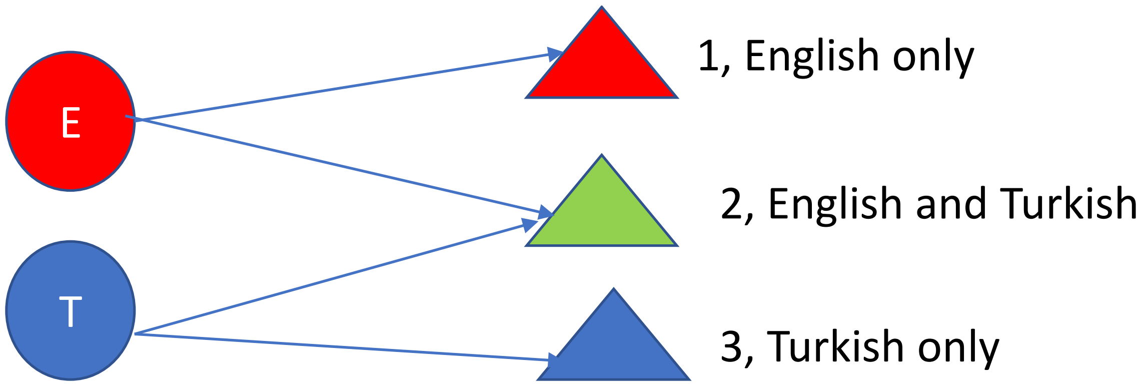
where X_E and X_t are the answered calls in English and in Turkish.



Flexible Capacity Design

- Consider a demand realization $d_{E,i}$ and $d_{T,i}$. We have:

$$
\begin{aligned}
x_{E,i} &= w_{E,1} + w_{E,2} && (1)\\
x_{T,i} &= w_{T,2} + w_{T,3} && (2)\\
x_{E,i} &\leq d_{E,i} && (3)\\
x_{T,i} &\leq d_{T,i} && (4)\\
w_{E,1} &\leq q_1 && (5)\\
w_{T,3} &\leq q_3 && (6)\\
w_{E,2} + w_{T,2} &\leq q_2 && (7)
\end{aligned}
$$



Where we are headed

- How should we solve such problems when there is data for **Y** ?
- How should we solve such problems when there are features **X** for **Y** (covariates)?
- What if the data includes time series?
- We'll see that there can be many potential features even based on the time series information. Can we handle many features efficiently?
- What if the number of potential features is much larger than the sample size (200 features and a sample size of 100)?

Where we are headed

- Some relevant and interesting problems are dynamic in nature
- Can we handle data-based dynamic optimization?
  - Approximate stochastic dynamic programming / reinforcement learning

A typical problem

- In practice (reality), the probability distribution of $\mathbf{Y}$ is not known with certainty but we may have some past observations on hand for $\mathbf{Y}$: $(\mathbf{y_1}, \mathbf{y_2}, \dots \mathbf{y_n})$.
- We may have observed demands of (28,43) at the two stores on day 1, (52, 25) on day 2 and so on.
- We then have options to "fit" a joint probability distribution using the observations or use the demand observations as scenarios that become inputs to the optimization problem.
  - For instance, we may fit a bivariate normal distribution to the data that specifies, the means, the standard deviations and the correlation.
  - A little more on this later.

Fitting a probability distribution

- Let us assume that we have an i.i.d sample of observations for **Y** (after some data transformations).
  - Obtaining and i.i.d. sample requires cleaning up many things in practice through data transformations.

- Eventually, we have something that may look like: y_1 = 24, y_2=35, y_3=11, y_4=48,…, y_n=55.
- Or : y_1 = 24.2, y_2=35.4, y_3=11.9, y_4=48.1,…, y_n=55.3.
- We may plot the histogram of the data and explore its shape (monotone, unimodal, multimodal, symmetrical, skewed).
- And take a guess for continuous or a discrete distribution to fit.

Fitting a probability distribution

- Let's assume we have a sample of iid demand observations $d_1, d_2, \ldots d_n$.
- We think that this sample might correspond to a Poisson r.v. with parameter λ:

$$p_D(x) = \frac{\lambda^x e^{-\lambda}}{x!} \quad x = 0, 1, 2, \ldots$$

- Since λ is not known, We look for the value of λ that makes the sample as likely as possible. This is an optimization problem:

$$\max_{\lambda} \Pi_{i=1}^{n} p_D(d_i, \lambda) = \Pi_{i=1}^{n} \frac{\lambda_i^{d} e^{-\lambda}}{d_i!}$$

This approach to find the optimal fit of the parameter through likelihood maximization is called Maximum Likelihood Estimation (MLE).

Fitting a probability distribution (MLE)

- The solution of the above problem:

$$\lambda^* = \arg\max_{\lambda} \Pi_{i=1}^{n} p_D(d_i, \lambda)$$

  corresponds to the value that maximizes the likelihood of the sample with respect to a given distribution.

  And is called the Maximum Likelihood Estimation (MLE) estimator.

- To solve the optimization problem, we take the logarithm of the likelihood function to convert the product to a sum.

Ex: Poisson (λ), sample $x_1, x_2 \dots x_n$

The likelihood function:

$$L(x_1, x_2 \dots x_n; \lambda) = \frac{\lambda^{x_1} e^{-\lambda}}{x_1!} \cdot \frac{\lambda^{x_2} e^{-\lambda}}{x_2!} \cdots \frac{\lambda^{x_n} e^{-\lambda}}{x_n!}$$

We take logs to convert the product to a sum

$$\ell(x_1, x_2 \dots x_n; \lambda) = \log L(x_1, x_2 \dots x_n; \lambda) \quad x_1 \log\lambda - \lambda - \log(x_1!) + x_2 \log\lambda - \lambda - \log(x_2!)$$
$$+ \dots + x_n \log\lambda - \lambda - \log(x_n!)$$

$$\frac{d\ell}{d\lambda} = \frac{\sum x_i}{\lambda} - n \Rightarrow \lambda^* = \frac{\sum x_i}{n}$$

Ex: Normal (μ, σ)

$$L(x_1, x_2 \dots x_n; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_1-\mu)^2}{2\sigma^2}} \dots \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_n-\mu)^2}{2\sigma^2}}$$

$$l(x_1, x_2 \dots x_n; \mu, \sigma) - \frac{n}{2}\log(2\pi) - n\log\sigma - \frac{\sum(x_i-\mu)^2}{2\sigma^2}$$

$$\frac{dl}{d\mu} = \frac{+2\sum(x_i-\mu)}{2\sigma^2} \Rightarrow \frac{\sum x_i - n\mu}{\sigma^2} = 0 \Rightarrow \mu^* = \frac{\sum x_i}{n}$$

$$\frac{dl}{d\sigma} = -\frac{n}{\sigma} + 2\frac{\sum(x_i-\mu)^2}{2\sigma^3} \Rightarrow \sigma^2 n = \sum(x_i-\mu)^2$$

$$\Rightarrow \sigma^{*2} = \frac{\sum(x_i-\mu)^2}{n}$$

Fitting a probability distribution

- We are able to 'optimally' estimate the parameters of different distributions (e.g. Poisson, Binomial, negative binomial etc.) given the data available.
- We can then measure the distance of the candidate distribution to the sample by several different approaches.
- The Kolmogorov-Smirnov goodness-of-fit test uses the squared distance in an interval. We separate the real line into K intervals and for each interval we compute e_k the expected number of observations that falls in the interval in the candidate distribution and also count o_k, the number of observations that fall in the same interval.
- The K-S statistic:

$$\sum_{k=1}^{K} \frac{(o_k - e_k)^2}{e_k}$$

has a χ^2 distribution which leads to a simple hypothesis test.

Fitting a probability distribution

- We then find the best fitting distribution among many candidates by comparing the values of the K-S statistic.
- Or do the same for a different distance metric (such as the Kullback-Liebler (KL divergence))

$$KL(f : g) = \int f(x) \log\left(\frac{f(x)}{g(x)}\right) dx$$

Reminder: estimators and properties

- A crucial issue in statistics is to infer population properties from a finite sample. An estimator is a quantity that can be computed from the sample for this purpose.
- We might be interested in estimating the mean μ of a population for which have an iid sample $x_1, x_2, ..., x_n$.
- The average of the sample $\bar{x}$ is an estimator.
- But there are other estimators than $\bar{x}$. x_1 is also an estimator, $(2x_1 + x_2)/3$ is another one.
- In fact, any $f(x_1, x_2, ..x_n)$ is a potential estimator.

Reminder: estimators and properties

- Let us note that sample based estimators are themselves random variables. Each time we draw a new random sample, we'll get a different value for our estimator.
- Unbiasedness: A desirable property for an estimator is that it does not have a systematic error on the average (in expectation). The sample mean $\bar{X}$ is an unbiased estimator of the population mean since:

$$E[\bar{X}] = \mu.$$

- Note that there are many unbiased estimators: X_1 and $(2X_1 + X_2)/3$ are also unbiased. Since:

$$E[X_1] = E[(2X_1 + X_2)/3] = \mu.$$

Reminder: estimators and properties

- Variance of the Estimator: Among unbiased estimators, it makes sense to prefer one with a lower variance.
- Assuming that our sample has variance σ^2:

$$Var[\bar{X}] = \frac{\sigma^2}{n}$$

- whereas for the other estimators:

$$Var[X_1] = \sigma^2 \text{ and } Var[(2X_1 + X_2)/3] = \frac{5\sigma^2}{9}.$$

- We will see that for demand forecasting there is a trade-off between responsiveness and low variance.