Info 2950, Lecture 13
13 Mar 2018

Prob Set 4: due Mon night 19 Mar

Wed Office hour (this week only) -> Fri 11-noon Gates 242

Thu 15 Mar: please bring laptop to class

Midterm: Thu Mar 22, 1:25-2:40, Statler 185
Midterm: Thu Mar 22, 1:25-2:40, Statler 185

- counting / probability
- bayes theorem / naive bayes classifier,
- normal and poisson distributions
- simple python
(open notebook / open computer, so you'll have everything available.)

Ideally if you've done a conscientious job and understand everything on the probsets, then doesn't require much concentrated studying, but useful to have an idea of where everything is to make it easier to find in real time during exam.

On probsets so far, ps1 1,4,5; ps2 1,2,3,4,5; ps3 1,2; ps4 1,2,3; are either adapted from or similar to past midterm problems.

Programming minimal at most, but bring charged laptop …

Enter answers on formatted sheet, front of exam book. Show enough to get partial credit if answer is incorrect. Can be turned in entirely on paper, there's also the option of uploading anything electronic to the course website during exam.
What about some arbitrary probability distribution?

mean=281.3
std=99.2
Recall the "Chebychev Bound":

For all distributions, and all numbers $k$, the proportion of entries that are in the range "average +/- $k$ SDs" is at least $1 - \frac{1}{k^2}$

This gives a lower bound, not an exact value or an approximation; but holds for all distributions, no matter how irregular, e.g.:

the proportion in the range "average +/- 2 SDs" is at least $1 - \frac{1}{4} = 0.75$
the proportion in the range "average +/- 3 SDs" is at least $1 - \frac{1}{9} = 0.89$
the proportion in the range "average +/- 4.5 SDs" is at least $1 - \frac{1}{4.5^2} = 0.95$

The percent of entries in the range "average +/- 2 SDs" might be much larger than 75%, as is the case for the normal distribution, but can never be smaller.

(check:
  for normal distribution, within +/- 2SD is 95%, clearly > 75% ;
  for normal distribution, within +/- 3SD is 99.7% > 89%)
(Normal distribution well above Chebyshev $1-1/k^2$ bound)
Chebyshev Bound: \( p(|x - z| \geq k\sigma) \leq 1/k^2 \) (where \( z = E[x] \))

Proof (for any distribution):

\[
\sigma^2 = E[(x - z)^2] = \sum_x (x - z)^2 p(x)
\]

\[
= \sum_{x \text{ s.t. } |x-z|<k\sigma} (x - z)^2 p(x) + \sum_{x \text{ s.t. } |x-z|\geq k\sigma} (x - z)^2 p(x)
\]

\[
\geq \sum_{x \text{ s.t. } |x-z|\geq k\sigma} (x - z)^2 p(x)
\]

\[
\geq \sum_{x \text{ s.t. } |x-z|\geq k\sigma} k^2 \sigma^2 p(x)
\]

\[
= k^2 \sigma^2 \sum_{x \text{ s.t. } |x-z|\geq k\sigma} p(x) = k^2 \sigma^2 p(|x - z| \geq k\sigma)
\]
Alternatively, use the Markov Inequality: If \( x \) takes only non-negative values, then \( p(x \geq a) \leq E[x]/a \)

\[
E[x] = \sum_{x < a} xp(x) + \sum_{x \geq a} xp(x)
\]

\[
\geq \sum_{x \geq a} xp(x)
\]

\[
\geq \sum_{x \geq a} ap(x) = a \sum_{x \geq a} p(x) = ap(x \geq a)
\]

Intuitive: can’t be much probability for \( x \geq a \) if \( a \) is much greater than \( E[x] \)

Then using the above Markov inequality with \( a = k^2\sigma^2 \):

\[
p[(x - z)^2 \geq k^2\sigma^2] \leq E[(x - z)^2]/k^2\sigma^2 = 1/k^2
\]

gives the Chebyshev Bound \( p(|x - z| \geq k) \leq 1/k^2 \)
**Definition:** A graph $G$ is a pair $(V, E)$, where $V$ is a finite set and $E$ is a collection of unordered pairs in $V \times V$.

The elements of $V$ are referred to as the vertices of the graph and the elements of $E$ are referred to as the edges.

**Example:** Define a graph $G_1$ with $V = \{1, 2, 3, 4, 5, 6\}$ and $E = \{(1, 2), (1, 3), (1, 4), (2, 4), (3, 4), (1, 5), (4, 6)\}$
We could have repeated elements in $E$.

For example, we could include an edge twice.

In this case, we say that $G$ has multiple edges.

We could also have an edge between just one vertex.

In this case, we call the edge a loop

Example: $(3, 3)$ is an edge and $(1, 5)$ has been included twice.

A simple graph is a graph with no loops and no multiple edges.
In our definition of a graph, edges are unordered pairs.

So the edge \((1, 4)\) is the same thing as the edge \((4, 1)\).

This kind of graph is referred to as **undirected**.

A **directed** graph is a graph whose edges are **ordered pairs** (shown by arrows in the points/lines representation).

For edge \((v_1, v_2)\), draw an arrow from \(v_1\) to \(v_2\).

**Example:** a directed graph with \(V = \{1, 2, 3, 4, 5, 6\}\) and 
\(E = \{(1, 2), (1, 3), (4, 1), (4, 2), (3, 4), (1, 5), (6, 4)\}\). 
A subgraph $H = (V', E')$ of a graph $G = (V, E)$ is a pair $V' \subseteq V$ and $E' \subseteq E$.

We say that $H$ is an induced subgraph of $G$ if all the edges between the vertices in $V'$ from $E$ are in $E'$.

**Example:** Two subgraphs of $G_1$. The first is an induced subgraph. All edges between the vertices 2, 3, 4, and 6 that are in $G_1$ are also in this graph. The second subgraph is not an induced subgraph because the edges (2, 4) and (1, 4) are missing.
We say that two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ are the same or isomorphic if there is a one to one and onto map $f$ from $V_1$ to $V_2$ such that $(f(v_i), f(v_j))$ is an edge in $G_2$ if and only if $(v_i, v_j)$ is an edge in $G_1$.

Example: The two graphs are isomorphic. In this example, we can simply take the map $f(1) = 1$, $f(2) = 2$, $f(3) = 3$, $f(4) = 5$, $f(5) = 2$, $f(6) = 6$.

In general, if graph has labeled vertices and we can find a map such that $f(i) = i$, then the graphs are isomorphic as labeled graphs.
We can change the labeling on one of the graphs so they’re no longer isomorphic as labeled graphs (they are still isomorphic as unlabeled graphs).

**Example:** In the labels 2 and 5 have been changed, but the graphs are still isomorphic.
Two **vertices** $v_1, v_2$ are adjacent or neighbors if $(v_1, v_2) \in E$. Two **edges** are adjacent if they share a common vertex.

A path from $v_1$ to $v_n$ in a graph is a sequence of adjacent edges such that $v_1$ is in the first edge of the sequence and $v_n$ is in the last edge of the sequence.

**Example:** the path $(5, 1)(1, 3)(3, 4)(4, 2)$ from 5 to 2 is highlighted.
We say a graph is connected if for every pair of vertices there exists a path between them. All the graphs so far have been connected. The figure below shows a graph that is not connected.

A disconnected graph consists of multiple connected pieces called components.
We call a vertex $v$ incident to an edge $e$ if $v \in e$.

The degree of a vertex $v$, written $\deg(v)$, is the number of edges to which it’s incident.

**Example:** In $G_1$, $\deg(1) = 4$, $\deg(2) = 2$, $\deg(3) = 2$, $\deg(4) = 4$, $\deg(5) = 1$, and $\deg(6) = 1$. 