Spatial Data Mining: Homework 3 (Due 11/11 at 11:59PM)

Use blue color to write your answers and submit on ELMS.

Writing Problems

1. (10 points) Concepts:

(1) Select all true statements about association rule mining.

A. Two events A and B with high association means A causes B or B causes A.
B. For a given itemset (e.g., ABC), support values of rules from the set may change.
C. For a given itemset (e.g., ABC), confidence values of rules from the set may change. True
D. For a given rule, support is an upper-bound of confidence, i.e., support ≥ confidence
E. For a given rule, support is a lower-bound of confidence, i.e., support ≤ confidence True

(2) Select all true statements about the apriori algorithm:

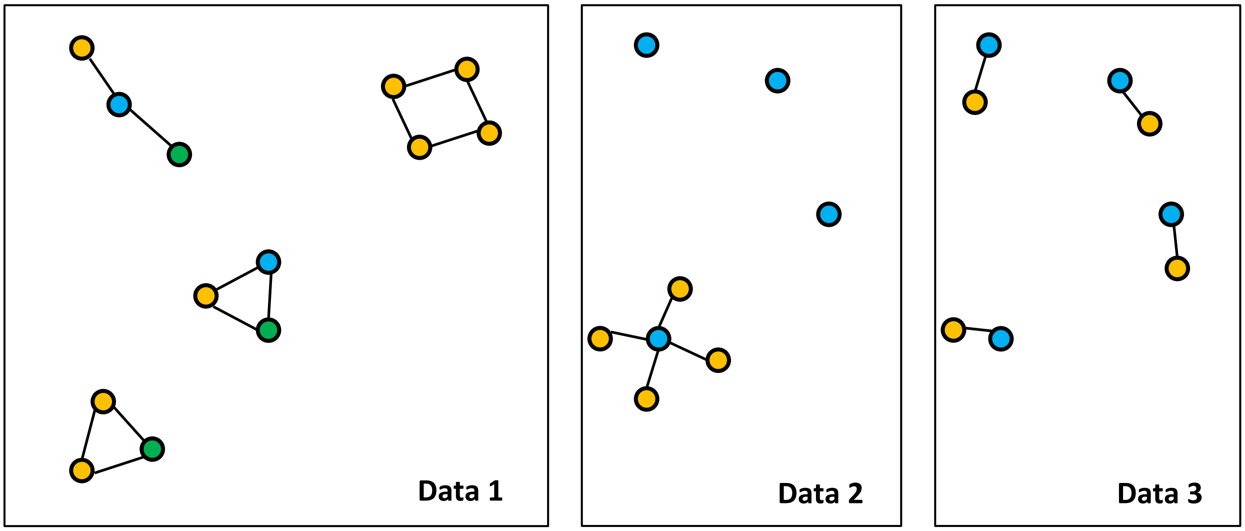
A. Apriori algorithm works by filtering out candidate patterns using support True
B. Apriori algorithm works by filtering out candidate patterns using confidence
C. Apriori algorithm accelerates the computation by sacrificing some solution quality
D. Apriori algorithm returns the same results as the brute-force (baseline) algorithm True
E. Support of an itemset is always smaller than or equal to the support of its subset True
F. Confidence of an itemset is always greater than or equal to the confidence of its subset

(3) Briefly explain if association between A and B is equivalent to causal relationships between A and B?

No, association between A and B is not always equal to a causal relationship between A and B because there could be a confounding variable. For instance, taking a look at the example of drinking an energy drink makes you fall asleep, but it could be because you're tired or didn't sleep enough the previous night.

2. (60 points) Calculation:

2.1 (30 points) Colocation (Spatial Association)



In datasets Data 1 to 3, the **neighbor relations are already given** (i.e., there is an edge between two points if and only if the distance between them is smaller than d). Event types are represented by point colors: Data 1 has three types, and Data 2 and Data 3 have two types each.

(1) **Data 1:** Calculate the **Participation Index** value for the following candidate patterns. Show important steps and quantities in your calculation.

Pattern 1: 🟡 🟢 i.e., (yellow, green)

Pr((yellow, green), yellow) = $\frac{3}{8}$

Pr((yellow, green), yellow) = $\frac{2}{3}$

Pi = $\min\{\frac{3}{8}, \frac{2}{3}\} = \frac{3}{8}$

Pattern 2: 🔵 🟢 i.e., (blue, green)

Pr((blue, green), blue) = $\frac{2}{2}$

Pr((blue, green), green) = $\frac{2}{3}$

Pi = $\min\{\frac{2}{2}, \frac{2}{3}\} = \frac{2}{3}$

Pattern 3: 🟡 🟢 🔵 i.e., (yellow, green, blue)

Pr((yellow, green, blue), yellow) = $\frac{1}{8}$

Pr((yellow, green, blue), green) = $\frac{1}{3}$

Pr((yellow, green, blue), blue) = $\frac{1}{2}$

Pi = $\min\{\frac{1}{8}, \frac{1}{3}, \frac{1}{2}\} = \frac{1}{8}$

(2) **Data 2:** Calculate the **Cross-K function** (suppose the area of the entire space to be 10, which is needed to calculate λ) and **Participation Index** values for the two event types.

$$\lambda_y = \lambda_b = \frac{4}{10} = \frac{2}{5}$$

$$k_y = \lambda^{-1}{}_y \times E\,(\#\ of\ yellow\ points\ within\ d\ of\ a\ blue\ point) = \frac{5}{2} \times \frac{4}{4} = \frac{5}{2}$$

$$k_b = \lambda^{-1}{}_b \ \times E\ (\#of\ blue\ points\ within\ d\ of\ a\ yellow\ point) = \frac{5}{2} \times \frac{1}{4} = \frac{5}{8}$$

$$pi(y,b) = \min\ \{\ pr\big((y,b),y\big), pr\big((y,b),b\big)\ \} = \min\left\{\frac{4}{4}, \frac{1}{4}\right\} = \frac{1}{4}$$

(3) **Data 3:** Calculate the **Cross-K function** (suppose the area of the entire space to be 10, which is needed to calculate λ) and **Participation Index** values for the two event types.

$$\lambda_y = \lambda_b = \frac{4}{10} = \frac{2}{5}$$

$$k_y = \lambda^{-1}{}_y \ \times E\ (\#\ of\ yellow\ points\ within\ d\ of\ a\ blue\ point) = \frac{5}{2} \times \frac{4}{4} = \frac{5}{2}$$

$$k_b = \lambda^{-1}{}_b \ \times E\ (\#of\ blue\ points\ within\ d\ of\ a\ yellow\ point) = \frac{5}{2} \times \frac{4}{4} = \frac{5}{2}$$

$$pi(y,b) = \min\ \{\ pr\big((y,b),y\big), pr\big((y,b),b\big)\ \} = \min\left\{\frac{4}{4}, \frac{4}{4}\right\} = 1$$

2.2 Matrix calculation basics for machine learning.

(1) (10 points) Write down the dimensions of the results for the following matrix multiplications.

Dimensions of matrices used for multiplication: $\boldsymbol{X} \in R^{N\times d}$, $\boldsymbol{y} \in R^{N\times 1}$, $\boldsymbol{w} \in R^{d\times 1}$

If a multiplication is NOT calculable, write "Incorrect Multiplication." Otherwise, show the result dimension ($Number\ of\ rows \times Number\ of\ columns$).

Multiplication	Result dimension ($Row \times Column$)
Example: $\boldsymbol{X}^T\boldsymbol{X}$	$d \times d$
$\boldsymbol{Xw}$	N x 1
$\boldsymbol{w}^T\boldsymbol{Xy}$	Incorrect multiplication
$\boldsymbol{XX}^T$	N x N
$\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{Xw}$	1 x 1

(2) (10 points) Show the gradient of the following function f(**w**) with respect to **w**. Hint: Result is a vector.

$$\boldsymbol{w} = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} \qquad f(\boldsymbol{w}) = (w_1)^2 + 2w_2$$

$$\frac{\partial f(\boldsymbol{w})}{\partial \boldsymbol{w}} = \begin{bmatrix} \frac{\partial f(\boldsymbol{w})}{\partial w_1} \\ \frac{\partial f(\boldsymbol{w})}{\partial w_2} \\ \frac{\partial f(\boldsymbol{w})}{\partial w_3} \end{bmatrix} = \begin{bmatrix} 2w_1 \\ 2 \\ 0 \end{bmatrix}$$

(3) **(Graduate Only,** 10 points**)** Show the gradient of the following function f(**w**) with respect to **w**. Show the steps and full resulting matrix. Hint: Feel free to use notations such as $\sum_{i=1}^{3} x_{1i} \cdot w_i$ to simplify your derivation (only an example).

$$\boldsymbol{w} = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} \qquad \boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \end{bmatrix} \quad \boldsymbol{e} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad f(\boldsymbol{w}) = \boldsymbol{e}^T \boldsymbol{X} \boldsymbol{w} - \boldsymbol{w}^T \boldsymbol{w}$$

$$\boldsymbol{e}^{\boldsymbol{T}} \boldsymbol{X} \boldsymbol{w} = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \end{bmatrix} \cdot \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} = \sum_{i=1}^{3} x_{1i} w_i \, , \sum_{i=1}^{3} x_{2i} w_i \rightarrow * \; [1 \quad 1] = \begin{bmatrix} \sum_{i=1}^{3} x_{1i} w_i \\ \sum_{i=1}^{3} x_{2i} w_i \end{bmatrix}$$

$$\boldsymbol{e}^{\boldsymbol{T}} \boldsymbol{X} \boldsymbol{w} = \sum\nolimits_{i=1}^{3} x_{1i} w_i + \sum\nolimits_{i=1}^{3} x_{2i} w_i$$

$$\boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{w} = [w_1 \quad w_2 \quad w_3] \cdot \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} = \boldsymbol{w_1^2} + \boldsymbol{w_2^2} + \boldsymbol{w_3^2} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{3}} \boldsymbol{w_i^2}$$

$$f(\boldsymbol{w}) = e^T X w - w^T w = \sum\nolimits_{i=1}^{3} x_{1i} w_i + \sum\nolimits_{i=1}^{3} x_{2i} w_i - \sum_{\boldsymbol{i=1}}^{\boldsymbol{3}} \boldsymbol{w_i^2}$$

$$\frac{\boldsymbol{\partial f(w)}}{\boldsymbol{\partial w_1}} = x_{11} + x_{21} - w_1$$

$$\frac{\boldsymbol{\partial f(w)}}{\boldsymbol{\partial w_2}} = x_{12} + x_{22} - w_2$$

$$\frac{\boldsymbol{\partial f(w)}}{\boldsymbol{\partial w_1}} = x_{13} + x_{23} - w_3$$

$$\frac{\boldsymbol{\partial f(w)}}{\boldsymbol{\partial w}} = \begin{bmatrix} x_{11} + x_{21} - w_1 \\ x_{12} + x_{22} - w_2 \\ x_{13} + x_{23} - w_3 \end{bmatrix}$$

3. (30 points for undergraduates; 20 points for graduates) Practice/Software

Here we will practice SaTScan, which is one of the most popular software for spatial scan statistics. It implements hotspot detection for many statistical models, including Bernoulli, Poisson and Normal distributions. In this question, you will use SaTScan to identify clusters using the Poisson model. The input will be case data only, and we will assume the underlying control distribution is homogeneous and continuous. Please see a step-by-step tutorial provided alongside the homework on how to use the software.

3.1 Run the software on the dataset **cluster.csv** (provided). Report significant clusters for **significance level $\boldsymbol{\alpha = 0.05}$**, and a study area that has $x_{min} = 0$, $y_{min} = 0$, $x_{max} = 100$, $y_{max} = 100$ (using "polygon inequalities" in the "polygon" window in the software, as shown in the step-by-step demo pdf). Include significant clusters' locations, radii, log-likelihood-ratios and p-values.

```
Location IDs included.: 998, 333, 269, 366, 414, 666, 959, 306, 526, 550, 236, 227, 295, 6,
                        471, 635, 258, 297, 298, 629, 761, 906, 696, 990, 466, 290, 682, 868,
                         669, 332, 237, 642, 282, 705, 874, 537, 620, 988, 838, 120, 818, 210,
                        577, 75, 254, 658, 259, 338, 495, 203, 285, 280, 503, 48, 508, 521,
                        712, 176, 39, 825, 923, 461, 776, 47, 213, 397, 981, 429, 160, 741,
                        889, 303, 650, 994, 467, 530, 94, 808, 742, 596, 525, 456, 598, 126,
                        52, 812, 108, 478, 360, 64, 136, 463, 805, 480, 276, 746, 485, 820,
                        328, 200, 95, 54, 289, 233, 980, 296, 390, 607, 524, 964, 36, 685,
                        749, 783, 38, 775, 501, 611, 728, 283, 49, 969, 726
  Coordinates / radius.: (37.74,42) / 11.54
  Observed / expected...: 2.94
  Log likelihood ratio.: 55.078587
  P-value...............: 0.001
```

3.2 **Increase the study area to $\boldsymbol{x_{min} = 0, y_{min} = 0, x_{max} = 1000, y_{max} = 1000}$** (using "polygon inequalities" in the "polygon" window in the software, as shown in the step-by-step demo pdf).

(1) Keep using the **cluster.csv** data, and report significant clusters for **significance level $\boldsymbol{\alpha = 0.05}$**.

```
Location IDs included.: 962, 625, 219, 364, 18, 139, 631, 753, 545, 413, 36, 232, 630, 549,
                 126, 181, 480, 271, 525, 20, 977, 746, 624, 160, 812, 508, 479, 461,
                 94, 727, 465, 923, 360, 934, 607, 963, 64, 213, 429, 969, 806, 75,
                 506, 667, 120, 647, 797, 696, 121, 108, 757, 620, 290, 117, 198, 789,
                 577, 330, 298, 258, 635, 644, 657, 981, 227, 6, 980, 236, 948, 555,
                 867, 283, 141, 414, 300, 666, 712, 551, 692, 550, 450, 259, 638, 295,
                 69, 521, 114, 166, 851, 485, 136, 579, 190, 49, 184, 390, 982, 868,
                 163, 912, 562, 471, 526, 958, 906, 619, 994, 937, 808, 765, 669, 893,
                 682, 998, 269, 321, 41, 615, 704, 890, 306, 805, 742, 530, 39, 440,
                 887, 123, 616, 889, 818, 53, 366, 130, 333, 332, 589, 482, 486, 466,
                 629, 787, 462, 927, 305, 761, 495, 512, 393, 959, 838, 939, 101, 598,
                 852, 233, 990, 90, 611, 363, 857, 617, 254, 873, 348, 216, 483, 372,
                 585, 415, 920, 749, 820, 289, 313, 297, 106, 510, 24, 375, 791, 328,
```

456, 487, 570, 935, 825, 282, 874, 713, 642, 143, 285, 58, 237, 710, 627, 280, 945, 275, 478, 975, 524, 210, 988, 192, 705, 534, 537, 538, 501, 54, 658, 854, 798, 356, 997, 338, 167, 467, 915, 596, 176, 503, 203, 48, 312, 741, 780, 47, 726, 680, 276, 95, 397, 783, 685, 776, 839, 170, 1, 34, 40, 382, 422, 744, 303, 418, 979, 432, 989, 650, 723, 463, 688, 618, 169, 52, 684, 572, 30, 443, 729, 964, 320, 882, 200, 960, 662, 81, 698, 886, 640, 97, 699, 686, 257, 7, 809, 296, 156, 728, 426, 33, 38, 777, 832, 404, 207, 875, 637, 708, 182, 226, 76, 933, 354, 846, 775, 474, 745, 991, 681, 513, 665, 974, 180, 79, 70, 301, 717, 716, 115, 605, 755, 540, 811, 309, 614, 660, 931, 102, 339, 118, 424, 909, 337, 264, 779, 781, 351, 400, 177, 491, 308, 944, 489, 738, 760, 255, 299, 689, 367, 904, 316, 919, 499, 697, 334, 187, 371, 691, 649, 505, 155, 687, 196, 721, 44, 604, 145, 888, 154, 536, 822, 567, 408, 84, 591, 10, 930, 307, 434, 764, 675, 622, 162, 671, 770, 938, 336, 772, 736, 386, 646, 581, 168, 310, 913, 816, 111, 2, 863, 803, 782, 468, 134, 318, 695, 199, 357, 14, 17, 752, 292, 830, 750, 864, 573, 211, 507, 15, 80, 693, 405, 878, 402, 870, 597, 531, 831, 853, 57, 718, 68, 950, 164, 580, 651, 389, 725, 807, 412, 186, 344, 700, 821, 758, 955, 801, 883, 381, 458, 490, 559, 151, 599, 473, 37, 261, 574, 146, 517, 59, 542, 802, 899, 444, 96, 209, 967, 986, 277, 92, 940, 74, 399, 901, 147, 268, 384, 137, 362, 720, 535, 42, 928, 672, 454, 437, 648, 743, 423, 668, 492, 826, 488, 553, 133, 217, 907, 563, 502, 476, 848, 349, 125, 655, 346, 112, 603, 225, 353, 785, 546, 419, 394, 189, 420, 442, 3, 674, 586, 105, 376, 85, 936, 272, 407, 417, 104, 113, 62, 144, 148, 358, 173, 824, 452, 869, 576, 733, 722, 25, 664, 387, 241, 447, 719, 428, 754, 377, 972, 127, 311, 262, 83, 11, 202, 350, 439, 996, 565, 234, 872, 843, 862, 766, 51, 245, 242, 961, 606, 288, 731, 77, 244, 520, 23, 71, 871, 449, 643, 877, 659, 107, 560, 578, 352, 568, 601, 759, 952, 60, 953, 884, 690, 158, 425, 652, 677, 433, 654, 946, 519, 543, 641, 73, 86, 529, 493, 518, 835, 315, 999, 564, 314, 441, 378, 110, 55, 66, 895, 179, 392, 532, 905, 477, 709, 294, 197, 435, 293, 132, 985, 21, 514, 342, 634, 916, 459, 956, 557, 149, 129, 922, 582, 445, 304, 971, 636, 188, 135, 183, 769, 547, 385, 284, 142, 157, 626, 511, 140, 175, 845, 898, 119, 902, 78, 46, 365, 361, 771, 706, 195, 398, 345, 287, 613, 22, 421, 894, 610, 26, 368, 623, 131, 796, 676, 247, 794, 246, 829, 496, 773, 100, 632, 833, 430, 849, 497, 5, 67, 800, 253, 50, 89, 281, 383, 774, 583, 355, 735, 756, 122, 335, 391, 243, 932, 327, 828, 663, 879, 850, 892, 460, 98, 347, 841, 56, 976, 124, 881, 325, 683, 855, 804, 302, 951, 941, 837, 661, 379, 817, 639, 410, 165, 896, 587, 516, 844, 575, 786, 947, 406, 152, 866, 359, 193, 924, 921, 91, 451, 31, 973, 380, 608, 859, 201, 541, 9, 317, 724, 594, 865, 87, 767, 267, 431, 214, 949, 515, 223, 171, 707, 438, 827, 343, 966, 810, 861, 205, 860, 965, 194, 455, 715, 880, 88, 714, 208, 265, 876, 178, 856, 509, 266, 694, 323, 970, 983, 763, 409, 730, 612, 504, 238, 4, 28, 656, 436, 252, 93, 993, 834, 369, 61, 472, 858, 734, 847, 204, 908, 911, 751, 222, 215, 815, 679, 185, 558, 279, 370, 249, 628, 799, 220, 229, 373, 99, 592,

```
                          174, 588, 212, 322, 673, 903, 778, 8, 554, 224, 470, 914, 995, 768,
                          653, 324, 446, 900, 16, 569, 274, 926, 836, 788, 602, 484, 475, 929,
                          138, 457, 739, 403, 1000, 978, 103, 291, 12, 150, 240, 556, 498, 82,
                          925, 13, 248, 251, 943, 703, 711, 842, 790, 270, 968, 984, 63, 795,
                          286, 43, 341, 823, 128, 411, 701, 221, 561, 153, 784, 228, 609, 159,
                          533, 396, 230, 27, 278, 702, 793, 464, 740, 218, 45, 992, 678, 566,
                          256, 191, 917, 813, 957, 732, 331, 235, 500, 116, 401, 340, 416, 481,
                          494, 819, 263, 918, 522, 792, 206, 273, 552, 427, 737, 448, 633, 590,
                          388, 172, 32, 374, 239, 329, 319, 29, 35, 747, 260, 72, 453, 469,
                          954, 645, 987, 161, 548, 621, 814, 670, 840, 109, 528, 231, 595, 897,
                          593, 65, 600, 891, 910, 942, 885
  Coordinates / radius.: (51.52,51.18) / 63.97
  Number of cases.......: 988
  Expected cases........: 12.85
  Observed / expected...: 76.86
  Log likelihood ratio.: 4236.987791
  P-value...............: 0.001
```

(2) Switch the dataset to **random.csv** (randomly generated dataset in range [0,100] for X axis and [0,100] for Y axis), and report significant clusters for **significance level** $\boldsymbol{\alpha = 0.05}$.

No significant clusters were detected for the significance level

(4) Briefly explain what you observe after the increase in study area and why.
When the study area was increased,
After increasing the study area, I noticed that the likelihood got increased, also the number of points/locations in the significant cluster increased, and the radius got larger, unlike the previous smaller cluster. Intuitively, increasing the study area causes an area that seems dispersed at first to seem clustered.