Problem 1

A. Mutual Information From Discrete Binning
Two of functions in the class-approved libraries make short work of this problem:

- **numpy.histogram2d(x,y,bins)** "Compute the bi-dimensional histogram of two data samples."
- **chi2_contingency(observed, correction, lambda_)** "Chi-square test of independence of variables in a contingency table"
  - Set lambda_="log-likelihood"
  - The G-test statistic is proportional to the Kullback-Leibler divergence, by a factor of $2N$
  - Convert from nats to bits
  - `MI[gene_a,gene_b] = 0.5 * g / discrete_expr.sum() / np.log(2)`

B. Equal Density Binning
The coded solution is in CalcMI.py.
C. Kernel Density Estimation
The coded solution is in CalcMI.py.
D. ROC plotting
The coded solution in included in plot.py
Prediction Evaluation
We are to evaluate several cases, and describe which gives the greatest AUROC. For the ones given, the 7-bin uniform density does best with an AUROC of .662 .

Figure 1: 7 bins, Uniform Size

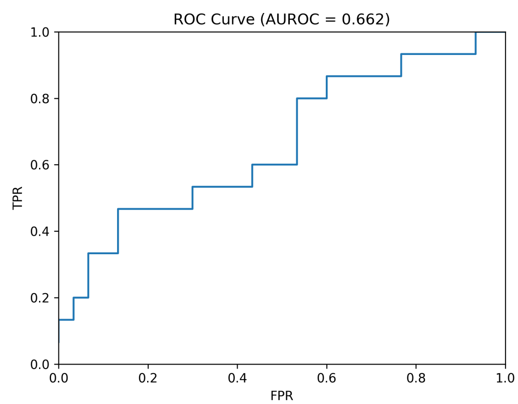


Figure 2: 9 bins, Equal Density

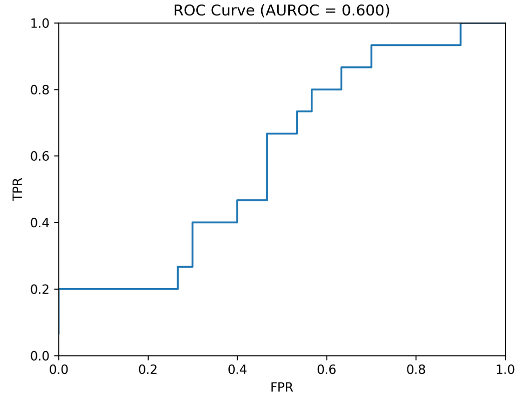
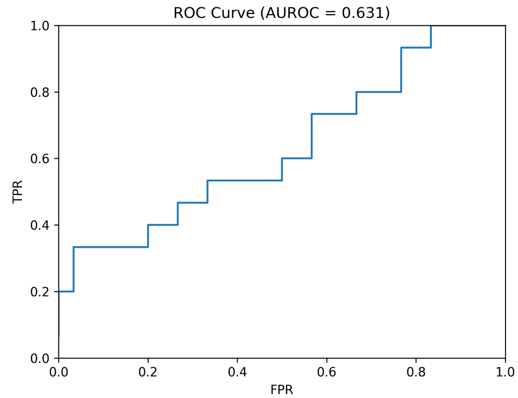


Figure 3: Kernel Estimatiom



Problem 2

2A Probabilities in the Markov random field

- How many possible configurations are there in this network? $2^{15} = 32768$
- What are the most probable and the least probable configurations?
  - The most probable configurations are when all nodes are in the same state: either on or off.
  - The least probable state is when nodes alternate between rows, so every pair is $(1, -1)$.
- Is it true that $P(R = 1|S = 1) = P(S = 1|R = 1)$?
  - No it is not. For the $P(S = 1|R = 1)$ case, we note that the calculation is unaffected by the state of the left branch of the tree, which is "shielded" by the given R state.
  - Because S is a leaf node, in the $P(R = 1|S = 1)$ case, the entire tree is still influencing R.
  - Therefore, $P(S = 1|R = 1) > P(R = 1|S = 1)$.

2B: Gibbs sampling in the Markov random field Here we program a gibbs sample, and estimate the probability of $P(R = 1|S = 1)$. Three iterations give values of .522, .528, and .547.
(To verify my answer to question 3 in part A, I estimated $P(S = 1|R = 1)$ to be about .55.)

Problem 3

3A Estimating λ Visually estimating λ is tricky because choosing the flattest part motivates $p > .6$:

$$\frac{4 * 1750}{.4 * 20000} \approx .85$$

However, I don't like this solution because the ideal curve is monotonically decreasing, if lambda was so high, there shouldn;t be so few p-values in the $(.2, .6)$ range. Therefore I would choose the bumpier ride of a threshold of $p = .2$.

$$\frac{8 * 1550}{.8 * 20000} \approx .78$$

In any case, as we are only asked to approximate to the nearest tenth, I'm rounding down to

$$\lambda = .8$$

3B Estimating $\hat{\pi}_0(\lambda)$

λ	$p_i > \lambda$	$\hat{\pi}_0(\lambda)$
0.0	20000	1.00
0.1	15427	0.86
0.2	12893	0.81
0.3	11382	0.81
0.4	9834	0.82
0.5	8466	0.85
0.6	7030	0.88
0.7	5259	0.88
0.8	3484	0.87
0.9	1714	0.86

3C: Calculating q-values

Rank	p-value	q-value
0	0.000003	0.048
1	0.000007	0.056
2	0.000013	0.069
3	0.000024	0.088
4	0.000028	0.088
5	0.000033	0.088
6	0.000046	0.105
7	0.000055	0.110
8	0.000096	0.158
9	0.000099	0.158