Probability and Stats Review

UW ECE 657A - Background Topic

Mark Crowley

Lecture Outline

Joint and Conditional Probability

Given event X (binary or multivalued). $p(X = x) = p(x)$ is the probability of the event that X takes on the value x.

Probability of A or B occurring:

$$\begin{aligned} p(A \vee B) &= p(A) + p(B) - p(A \wedge B) \\ &= p(A) + p(B) \text{ if A and B are mutually exclusive} \end{aligned}$$

Joint Probabilities: Product Rule and Chain Rule

$$p(A, B) = p(A \wedge B) = p(A|B)p(B) = p(B|A)p(A)$$

$$p(X_1, X_2, \ldots, X_D) = p(X_1)p(X_2|X_1)p(X_3|X_2, X_1)\ldots p(X_D|X_{1:D-1})$$

Marginal and Conditional Probability

Marginal Distribution:

$$p(A) = \sum_b p(A, B) = \sum_b p(A|B = b)p(B = b)$$

Conditional Probability:

$$p(A|B) = \frac{p(A, B)}{p(B)} \text{ if } p(B) > 0$$

"Probability of A given B"

Unconditional Independence

If two random variables variable X and Y are independent we denote it as $X \perp Y$

$$X \perp Y \text{ iff } p(X, Y) = p(X)p(Y)$$

Bayes Theorem

Given a hypothesis h and observed evidence e:

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

$$p(h|e) = \frac{p(e|h)p(h)}{p(e)}$$

$$p(cancer|testresult) = \frac{p(testresult|cancer)p(cancer)}{p(testresult)}$$

- *An aside:* Bayesian Statistics vs Frequentist Statistics
- very important, for knowing how to update a model based on new evidence, also tells you how to turn around a $p(X|Y)$ into a $p(Y|X)$

Bayes Theorem For Multidimensional Data

For data x with prediction target y:

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

$$p(y|x_1, \ldots, x_n) = \frac{p(x_1, \ldots, x_n|y)p(y)}{p(x_1, \ldots, x_n)}$$

If we knew that all of the features x_i were **independent** then we'd have:

$$p(y|x_1, \ldots, x_n) = \frac{p(y) \prod_{i=1}^{n} p(x_i|y)}{p(x_1, \ldots, x_n)}$$

Bayes Theorem as a Proportion

The probability of the evidence is constant and just for normalizing to a probability. So if we only want to compare probabilities we can drop it:

$$p(y|x_1, \ldots, x_n) = \frac{p(y)\prod_{i=1}^{n} p(x_i|y)}{p(x_1, \ldots, x_n)}$$

$$p(y|x_1, \ldots, x_n) \propto p(y)\prod_{i=1}^{n} p(x_i|y)$$

Entropy

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

- The higher the entropy the higher the uncertainty for that value.
- Also measures surprise of seeing the observation.
- How much information is represented by this observation.

Visualizing Entropy

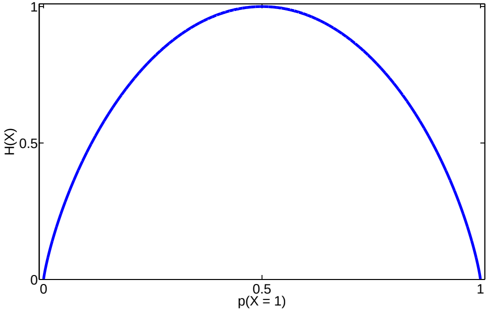

Figure: Binary Entropy Function: Entropy of the Bernoulli random variable as a function of θ. The maximum entropy is $\log_2 2 = 1$ when $\theta = 0.5$ (i.e. when the distribution is uniform).

Probabilistic Distance Methods

Earlier we talked about **Distance Measures** between *sets of datapoints*.
Similarly,

- if we know the distibution of our data
- or if we can infer it or learn it
- then we can measure the similarity (or equivalently, the *distance*) between the two underlying distributions.

Probabilistic Distance Methods

Earlier we talked about **Distance Measures** between *sets of datapoints*.
Similarly,

- if we know the distibution of our data
- or if we can infer it or learn it
- then we can measure the similarity (or equivalently, the *distance*) between the two underlying distributions.

Stop and Think

What does the distance between distributions look like?
What does it mean?

KL-Divergence or Relative Entropy

Kullback-Leibler Divergence (KL-Divergence) is a common method for measuring the dissimilarity between two probability distributions P and Q.
It can also be seen as a **Relative Entropy** measure between two distibutions.

$$KL(P||Q) = \sum_{i=1}^{N} P(i) \log \frac{P(i)}{Q(i)}$$

- $KL(P||Q) \geq 0$ and equals zero iff $P = Q$
- How much information you'd lose approximating Q with P
- In general $KL(P||Q) \neq KL(Q||P)$

Mutual Information (MI)

The **mutual information (MI)** between two vectors X, Y measures how similar the joint distribution $p(X, Y)$ is to the factored distribution $p(X)p(Y)$:

$$MI(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}$$

- MI(X,Y) is always nonnegative
- Equals 0 iff X, Y are independent
- Notice this is just the KL-Divergence between the distributions $p(X, Y)$ and $p(X)p(Y)$

Relation of MI to Entropy

The entropy $H(X)$ and MI are related:

$$H(X) = -\sum_{x \in X} p(x) \log p(x) \tag{1}$$

$$MI(X, Y) = H(X) + H(Y) - H(X, Y) \tag{2}$$

- MI can seen as the *reduction in entropy* on the labels that results from observing feature value x_j
- Some measures use MI normalized by the entropy $H(X)$

Information Gain Measure

Another measure you could use is **information gain**.

$$IG(Y, X) = H(Y) - H(Y|X)$$

Random Variables

See
https://rateldajer.github.io/ECE493T25S19/preliminaries/probabilityreview/

Probability Functions

Nice Distribution you have there...

but I gotta ask...is it significant?

Nice Distribution you have there...

but I gotta ask...is it significant?

- SECRET: Processing data and making predictions isn't our end goal...our goal is *understanding* or *interpretting* data and having *confidence* that our analysis is correct...or at least *plausible*.

What is Hypothesis Testing?

- Given a known distribution D_0 we think produced the data, call this our **null hypothesis** (often denoted H_0)
- Want to ask whether we can *reject the null hypothesis* given some observed data.
- Say D_0 is $N(0, 1)$ a standardized Gaussian and the sample is $x = 2.576$.
- $p(|u| \leq 2.576) = .99$: The probability of a sample u taken from $\mathcal{N}(0, 1)$ being less then 2.576 is 99%.
- So we say the difference of the sample x from the assumed distribution is **statistically significant**

So How Do We Use This?

- **Alternative Hypothesis**: What the scientist believes to be true.
- **Null Hypothesis**: What if the scientist is wrong?
- Goal: to say that the sample x lets us "reject the null hypothesis at the 0.1 confidence level".

So How Do We Use This?

- **Alternative Hypothesis**: What the scientist believes to be true.
- **Null Hypothesis**: What if the scientist is wrong?
- Goal: to say that the sample x lets us "reject the null hypothesis at the 0.1 confidence level".

So How Do We Use This?

- **Alternative Hypothesis**: What the scientist believes to be true.
- **Null Hypothesis**: What if the scientist is wrong?
- Goal: to say that the sample x lets us "reject the null hypothesis at the 0.1 confidence level".

Ways of Specifying the Null Hypothesis

There are many methods for doing this:

- for *discrete data* one is the **Chi-squared (χ^2) Test**.
- for a *Gaussian null hypothesis* you can use the **Student's t-Test**

Chi-squared (χ^2) Test

χ^2 statistics can be used to test whether a feature is statistically significant in predicting a class. For a feature x_f and class y_k we can formulate the Chi-square test:

$$\chi^2(x_{fi}, y_k) = \sum_{x_{fi} \in X_f} \sum_{y_k \in Y} \frac{(O_{ik} - E_{ik})^2}{E_{ik}}$$

$$= N \sum_{x_{fi} \in X_f} \sum_{y_k \in Y} p_i p_k \left(\frac{(O_{ik}/N) - p_i p_k}{p_i p_k} \right)^2$$

- O are the observed counts of joint events and E are their expected counts.
- χ^2 tests the hypothesis that;
  - the features and the classes are assigned randomly from a Gaussian distribution
  - and sampled i.i.d. from each class
- The higher the value of χ^2, the more likley we reject the null hypothesis of independent, random assignment of classes.

Contingency Table

The counts for χ^2 can be obtained using a contingency table

		y_k	
		1	0
x_{fi}	1	o_{11}	o_{12}
	0	o_{21}	o_{22}

- o_{11} is number of samples in the class that has the feature
- o_{21} is number of samples in the class that doesn't have the feature
- o_{12} is number of samples in other classes that has the feature
- o_{22} is number of samples in other classes that doesn't have the feature

Contingency Table

x_{fi} \ y_k	1	0	
1	100	70	= 170
0	60	30	= 90
	= 160	= 100	N = 260

E_{ik}	1	0
1	104.6	65.4
0	55.4	34.6

$$E_{11} = (o_{11} + o_{21})(o_{11} + o_{12})/N \qquad E_{12} = (o_{12} + o_{22})(o_{11} + o_{12})/N \tag{3}$$

$$E_{21} = (o_{11} + o_{21})(o_{21} + o_{22})/N \qquad E_{22} = (o_{12} + o_{22})(o_{21} + o_{22})/N \tag{4}$$

Contingency Table

x_{fi} \ y_k	1	0	
1	100	70	= 170
0	60	30	= 90
	= 160	= 100	N = 260

E_{ik}	1	0
1	104.6	65.4
0	55.4	34.6

$$\chi^2 = \frac{(100-104.6)^2}{104.6} + \frac{(70-65.4)^2}{65.4} + \frac{(60-55.4)^2}{55.4} + \frac{(30-34.6)^2}{34.6} = 1.51 \tag{5}$$

The number of *degrees of freedom* here is 1. Now we can use a **Chi-squared lookup table** to find the critical value for this number at a desired significance level. We see that for p=0.05 we need $\chi^2 > 3.8$ to reject the null hypothesis and claim that our feature is significant. In this case we can only claim p=0.30 significance.

χ^2 Lookup Table

Degrees of freedom (df)	χ^2 value[18]										
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.87	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
P value (Probability)	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001

Problems/Challenges with Hypothesis Testing

- Null hypothesis might be very artificial and difficulty to specify.
- Confusing terminology and counterintuitive concepts make it easy to use Hypothesis Testing incorrectly.
- HT is not a replacement for controlled causal studies!
- HT does not tell you "how significant" your results are. It tells you...
- As number of data samples gets larger, or precision of measurements increases, the confidence goes up!
  - This is an important, but not a fatal, flaw
  - The Standard i.i.d. assumption in ML is almost equivalent and just as flawed.