Graduate Computer Vision

CS670

Unit 2: Probability, Statistics, Supervised Learning: Feature Selection

Erik Learned-Miller

Today

- Information theoretic quantities: entropy, joint entropy, KL-divergence, and mutual information
- Conditional entropy
- Conditional mutual information and information gain
- Optimal and greedy algorithms for feature selection

Notation

Entropy

- How much "information" do you get when you observe a random variable?
  - How many bits, *on average*, do you have to send to communicate the outcome of the random variable to someone else?
  - coin with probability distribution [½, ½]: 1 bit
- Definition:

The *entropy* of a discrete random variable X with probability distribution given by $P(X)$ is

$$H(X) = -\sum_{x \in \mathcal{X}} P(x) \log P(x).$$

Don't forget the negative sign!

Entropy

- A 4-sided die that always lands on side A or side C: [½ 0 ½ 0]

$$H(X) = -\left[\frac{1}{2}\log(\frac{1}{2}) + 0\log 0 + \frac{1}{2}\log(\frac{1}{2}) + 0\log 0\right]$$

- How should we evaluate 0 log 0?

Entropy

- A 4-sided die that always lands on side A or side C: [½ 0 ½ 0]

$$H(X) = -\left[\frac{1}{2}\log(\frac{1}{2}) + 0\log 0 + \frac{1}{2}\log(\frac{1}{2}) + 0\log 0\right]$$

- How should we evaluate 0 log 0?

$$\lim_{x\to 0} P(x)\log P(x) = 0.$$

Why does this make sense?

A coin can land on its edge…

True or False?

- The entropy of independent random variables is the sum of the entropies of each variable?

Entropy of Ind. RVs

$$
\begin{aligned}
& H(X,Y) & (1)\\
= & -\sum_{(x,y)\in(\mathcal{X},\mathcal{Y})} P(x,y)\log P(x,y) & (2)\\
= & -\sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} P(x,y)\log P(x,y) & (3)\\
= & -\sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} P(x)P(y)\log P(x)P(y) & (4)\\
= & -\sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} P(x)P(y)[\log P(x)+\log P(y)] & (5)\\
= & -\sum_{y\in\mathcal{Y}} P(y)\log P(y)\sum_{x\in\mathcal{X}} P(x) - \sum_{y\in\mathcal{Y}} P(y)\sum_{x\in\mathcal{X}} P(x)\log P(x) & (6)\\
= & H(X)+H(Y). & (7)
\end{aligned}
$$

Joint Entropy

- The joint entropy of P(X,Y) is just the same as the entropy of a single random variable Z, where Z is a renaming of (X,Y):
  - Example: entropy of (precipitation, temperature) vs. entropy of “weather”.

KL-divergence (relative entropy)

- How different are two probability distributions?

$$D(P(X)\|Q(X)) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}$$

KL-divergence (relative entropy)

- How different are two probability distributions?

$$D(P(x)\|Q(x)) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}$$

What happens when P(x) = Q(x)?

KL-divergence (relative entropy)

- How different are two probability distributions?

$$D(P(x)\|Q(x)) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}$$

What happens when P(x) = Q(x)?

- Not symmetric: Order of P, Q matters!

Mutual Information

$$I(X;C) = \sum_{x \in \mathcal{X}, c \in \mathcal{C}} P(x,c) \log \frac{P(x,c)}{P(x)P(c)}$$

Feature Selection

- If we can have only one feature, which feature should we have?

Feature Selection

- If we can have only one feature, which feature should we have?
  - Feature whose mutual information with class label is highest.

Feature Selection

- Let possible features be called $X_1, X_2, ..., X_k$

  We would like

$$
\begin{aligned}
& \arg\max_{1 \leq i \leq k} I(X_i; C) \\
= \quad & \arg\max_{1 \leq i \leq k} \sum_{x_i \in \mathcal{X}_i, c \in \mathcal{C}} P(x_i, c) \log \frac{P(x_i, c)}{P(x_i) P(c)}
\end{aligned}
$$

Best 2 features

- Let possible features be called $X_1, X_2, ..., X_k$

  We would like

$$
\begin{aligned}
& \underset{1 \le i,j \le k}{\arg\max}\, I(X_i, X_j; C) \\
= \; & \underset{1 \le i,j \le k}{\arg\max} \sum_{(x_i, x_j) \in (\mathcal{X}_i, \mathcal{X}_j), c \in \mathcal{C}} P(x_i, x_j, c) \log \frac{P(x_i, x_j, c)}{P(x_i, x_j) P(c)}
\end{aligned}
$$

Best 3 features... yikes

- Let possible features be called $X_1, X_2, ..., X_k$

  We would like

$$
\begin{aligned}
& \operatorname*{arg\,max}_{1 \le i,j,h \le k} I(X_i, X_j, X_h; C) \\
= \; & \operatorname*{arg\,max}_{1 \le i,j,h \le k} \sum_{(x_i, x_j, x_h) \in (\mathcal{X}_i, \mathcal{X}_j, \mathcal{X}_h), c \in \mathcal{C}} P(x_i, x_j, x_h, c) \log \frac{P(x_i, x_j, x_h, c)}{P(x_i, x_j, x_h) P(c)}
\end{aligned}
$$

Let's analyze this

- There are 2 problems with finding the optimal set of features:
  - Computational complexity (obvious)
  - Statistical complexity (subtle)

Computational Complexity

- k features
- Number of features to try:
  - Best feature: $O(k)$
  - Best two features: $O(k^2)$
  - Best three features: $O(k^3)$
- Computing mutual information for each choice:
  - 1 feature: $O(|X|)$
  - 2 features: $O(|X|^2)$
  - 3 features: $O(|X|^3)$

Statistical Complexity

- How many samples of each joint distribution do we need to ensure confidence in our results?

$$P(x_i)$$

$$P(x_i, x_j)$$

$$P(x_i, x_j, x_h)$$

Greedy Approach

- First find best single feature.
- Then find best second feature given first.
- Find 3rd feature given the first two.

Best feature

- Let possible features be called $X_1, X_2, ..., X_k$

  We still want.

$$\begin{aligned} & \arg\max_{1 \le i \le k} I(X_i; C) \\ = \quad & \arg\max_{1 \le i \le k} \sum_{x_i \in \mathcal{X}_i, c \in \mathcal{C}} P(x_i, c) \log \frac{P(x_i, c)}{P(x_i)P(c)} \end{aligned}$$

Best 2nd feature given first.

- Let possible features be called $X_1, X_2, ..., X_k$

$$\arg\max_{1 \le j \le k} I(X_i, X_j; C) - I(X_i; C).$$

Best 2nd feature given first.

- Let possible features be called $X_1, X_2, ..., X_k$

$$\arg\max_{1 \leq j \leq k} I(X_i, X_j; C) - I(X_i; C).$$

maximize the *information gain*.

More than 2 features

- How do we handle the best 3 features?
  best 4 features?
  best 5 features?

More than 2 features

- Assume we already have chosen f features. Which feature from among k should we choose next?

Example

- Assume we have already chosen as features X_A, X_B
- We will NOT compute this:

$$\arg\max_{1 \leq i \leq k} I(X_i, X_A, X_B; C) - I(X_A, X_B; C).$$

Why not?

Alternative

- Try to make sure that new feature is not "highly redundant" with any previous feature.
  Consider two possible new features X_C and X_D

$$I(X_A, X_C; C) - I(X_A; C) = 0.3$$
$$I(X_B, X_C; C) - I(X_B; C) = 0$$

$$I(X_A, X_D; C) - I(X_A; C) = 0.2$$
$$I(X_B, X_D; C) - I(X_B; C) = 0.1$$

How much value do we get from adding X_C ?

Final Greedy Strategy for Nth feature

$$\arg\max_{1 \le i \le k} \left[\min_{F \in \mathcal{F}} I(X_i, X_F; C) - I(X_F; C) \right].$$

Means

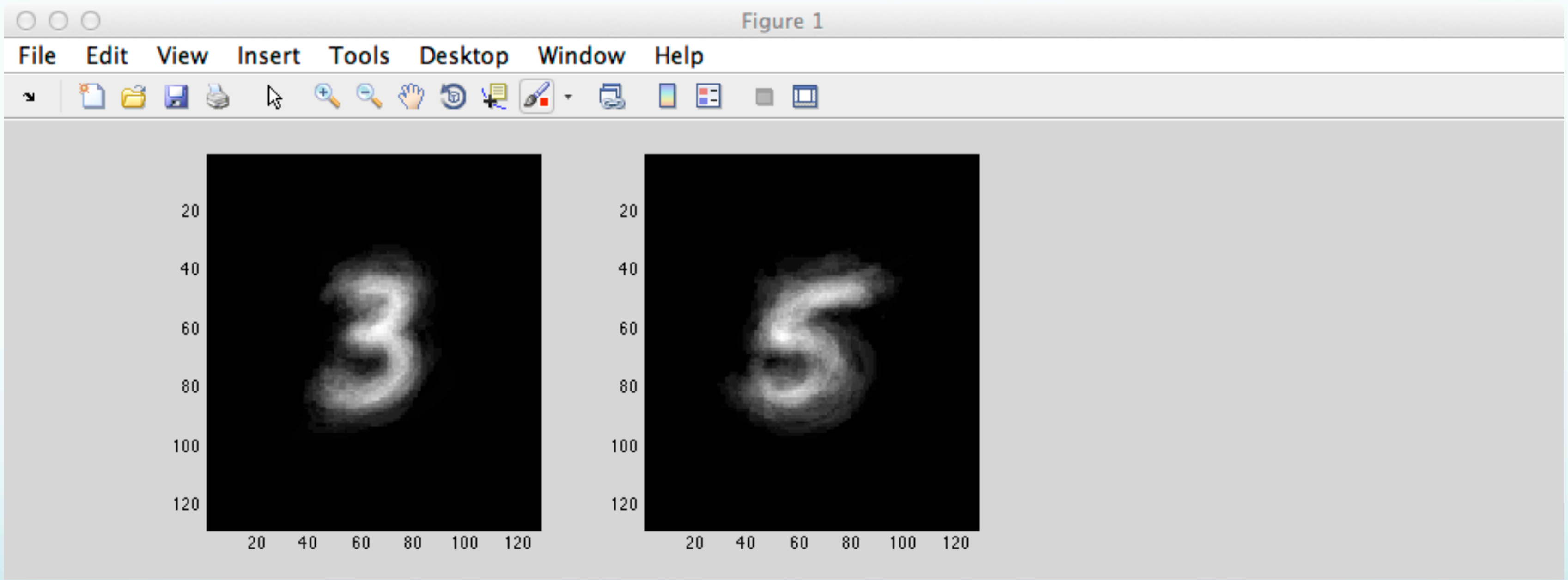
Figure 1
File Edit View Insert Tools Desktop Window Help
20
40
60
80
100
120
20 40 60 80 100 120
20
40
60
80
100
120
20 40 60 80 100 120

Means and differences of means