CMPSCI 240: "Reasoning Under Uncertainty"

Lecture 19

Not-A-Prof. Phil Kirlin
pkirlin@cs.umass.edu

April 3, 2012

Recap

Hypothesis Testing

- Let D be the event that we have observed some data, e.g., D = observed an email containing "ca$h" and "viagra"

Hypothesis Testing

- Let D be the event that we have observed some data, e.g., D = observed an email containing "ca$h" and "viagra"
- Let $H_1, \ldots, H_k$ be disjoint, exhaustive events representing hypotheses that we want to choose between, e.g., H_1 = event that email is spam, H_2 = event that email is not spam

Hypothesis Testing

- Let D be the event that we have observed some data, e.g., D = observed an email containing "ca$h" and "viagra"
- Let $H_1, \ldots, H_k$ be disjoint, exhaustive events representing hypotheses that we want to choose between, e.g., H_1 = event that email is spam, H_2 = event that email is not spam
- How do we use D to decide which hypothesis is most likely?

Bayesian Reasoning (Recap)

- If we have k disjoint, exhaustive hypotheses $H_1, \ldots, H_k$ (e.g., spam, not spam) and some observed data D (e.g., certain words in an email), we can use Bayes' theorem to compute the conditional probability $P(H_i \mid D)$ of hypothesis H_i $(i = 1, \ldots, k)$ given D:

Bayesian Reasoning (Recap)

- If we have k disjoint, exhaustive hypotheses $H_1, \ldots, H_k$ (e.g., spam, not spam) and some observed data D (e.g., certain words in an email), we can use Bayes' theorem to compute the conditional probability $P(H_i \mid D)$ of hypothesis H_i $(i = 1, \ldots, k)$ given D:

  $$P(H_i \mid D) = \frac{P(D \mid H_i)P(H_i)}{P(D)}$$

  where

Bayesian Reasoning (Recap)

- If we have k disjoint, exhaustive hypotheses $H_1, \ldots, H_k$ (e.g., spam, not spam) and some observed data D (e.g., certain words in an email), we can use Bayes' theorem to compute the conditional probability $P(H_i \mid D)$ of hypothesis H_i ($i = 1, \ldots, k$) given D:

$$P(H_i \mid D) = \frac{P(D \mid H_i)P(H_i)}{P(D)}$$

where

$$P(D) = \sum_{i=1}^{k} P(H_i)P(D \mid H_i)$$

Choosing the "Best" Hypothesis (Recap)

- Sometimes we have all those pieces of information, sometimes we don't.
- There are two ways to pick the "best" hypothesis, depending on what information we have available.

Maximum Likelihood (Recap)

Definition

The maximum likelihood hypothesis H^{ML} for observed data D is the hypothesis H_i $(i = 1, \ldots, k)$ that maximizes the likelihood:

$$H^{\text{ML}} = \underset{i}{\operatorname{argmax}} \; P(D \mid H_i)$$

Maximum Likelihood (Recap)

Definition

The maximum likelihood hypothesis H^{ML} for observed data D is the hypothesis H_i $(i = 1, \ldots, k)$ that maximizes the likelihood:

$$H^{\mathrm{ML}} = \operatorname*{argmax}_i \; P(D \mid H_i)$$

The maximum likelihood hypothesis H^{ML} is the hypothesis that assigns the highest probability to the observed data D

Maximum Likelihood (Recap)

Definition

The maximum likelihood hypothesis H^{ML} for observed data D is the hypothesis H_i $(i = 1, \ldots, k)$ that maximizes the likelihood:

$$H^{\mathrm{ML}} = \underset{i}{\operatorname{argmax}}\; P(D \mid H_i)$$

The maximum likelihood hypothesis H^{ML} is the hypothesis that assigns the highest probability to the observed data D

How to use it: compute the $P(D \mid H_i)$ for all $i = 1, \ldots, k$ hypotheses and then select the hypothesis with the greatest value

Maximum A Posteriori (MAP) Hypothesis

Definition

The MAP hypothesis H^{MAP} for observed data D is the hypothesis H_i $(i = 1, \ldots, k)$ that maximizes the posterior probability:

$$\begin{aligned} H^{\mathsf{MAP}} &= \operatorname*{argmax}_i \; P(H_i \mid D) \\ &= \operatorname*{argmax}_i \; \frac{P(D \mid H_i) P(H_i)}{P(D)} \\ &\propto \operatorname*{argmax}_i \; P(D \mid H_i) P(H_i) \end{aligned}$$

Maximum A Posteriori (MAP) Hypothesis

Definition

The MAP hypothesis H^{MAP} for observed data D is the hypothesis H_i $(i = 1, \ldots, k)$ that maximizes the posterior probability:

$$
\begin{aligned}
H^{\text{MAP}} &= \operatorname*{argmax}_i \; P(H_i \mid D) \\
&= \operatorname*{argmax}_i \; \frac{P(D \mid H_i) P(H_i)}{P(D)} \\
&\propto \operatorname*{argmax}_i \; P(D \mid H_i) P(H_i)
\end{aligned}
$$

The likelihoods are now *weighted* by the prior probabilities; unlikely hypotheses are therefore downweighted accordingly.

One Slide To Rule Them All

- The maximum likelihood hypothesis is the hypothesis that assigns the highest probability to the observed data:

$$H^{\mathsf{ML}} = \operatorname*{argmax}_i P(D \mid H_i)$$

- The maximum a posteriori (MAP) hypothesis is the hypothesis that that maximizes the posterior probability given D:

$$\begin{aligned} H^{\mathsf{MAP}} &= \operatorname*{argmax}_i P(H_i \mid D) \\ &= \operatorname*{argmax}_i \frac{P(D \mid H_i)\, P(H_i)}{P(D)} \\ &\propto \operatorname*{argmax}_i P(D \mid H_i)\, P(H_i) \end{aligned}$$

- $P(H_i)$ is called the prior probability (or just prior).
- $P(H_i|D)$ is called the posterior probability.

Example

A patient comes to visit Dr. Gregory House because they have a cough. After insulting and belittling the patient, House consults with his team of diagnosticians, who tell him that if a patient has a cold, then there's a 75% chance they will have a cough. But if a patient has the Ebola virus, there's a 80% chance they will have a cough.

What is the maximum likelihood hypothesis for the diagnosis?

Example

After concluding the patient has Ebola, House fires all his diagnosticians for their poor hypothesis testing skills and hires new ones. This new team does some background research and discovers if they are only going to consider the common cold and Ebola, then before the symptoms are even considered, there's a 1% chance the patient has Ebola and a 99% chance they have a cold.

What is the MAP hypothesis for the diagnosis? What is the posterior probability the patient has Ebola?

Combining Evidence Example

Suppose you're a CS grad student and therefore work in a windowless office. You want to know whether it's raining outside. The chance of rain is 70%. Your advisor walks in wearing his raincoat. If it's raining, there's a 65% chance he'll be wearing a raincoat. Since he's very unfashionable, there's a 45% chance he'll be wearing his raincoat even if it's not raining. Your officemate walks in with wet hair. When it's raining there's a 90% chance her hair will be wet. However, since she sometimes goes to the gym before work, there's a 40% chance her hair will be wet even if it's not raining. What's the posterior probability that it's raining?

Combining Evidence

- We can't solve this problem because we don't have any information about the probability of your advisor wearing a raincoat and your colleague having wet hair occurring simultaneously.

Combining Evidence

- We can't solve this problem because we don't have any information about the probability of your advisor wearing a raincoat and your colleague having wet hair occurring simultaneously.
- However, it is reasonable to assume that once we know whether it is raining or not, those events are conditionally independent of each other.

Combining Evidence

- We can't solve this problem because we don't have any information about the probability of your advisor wearing a raincoat and your colleague having wet hair occurring simultaneously.
- However, it is reasonable to assume that once we know whether it is raining or not, those events are conditionally independent of each other.
- This means $P(C \cap W \mid R) = P(C \mid R) \cdot P(W \mid R)$ (and similarly for the complementary event combinations).

Combining Evidence: Conditionally Independent Evidence

Definition

If we have k disjoint, exhaustive hypotheses $H_1, \ldots, H_k$ (e.g., rainy, dry) and m pieces of observed data that are conditionally independent given a hypothesis $D_1, \ldots, D_m$, then the posterior probability $P(H_i \mid D_1 \cap \ldots \cap D_m)$ of hypothesis H_i $(i = 1, \ldots, k)$ given the observed data $D_1 \cap \ldots \cap D_m$ is:

Combining Evidence: Conditionally Independent Evidence

Definition

If we have k disjoint, exhaustive hypotheses $H_1, \ldots, H_k$ (e.g., rainy, dry) and m pieces of observed data that are conditionally independent given a hypothesis $D_1, \ldots, D_m$, then the posterior probability $P(H_i \mid D_1 \cap \ldots \cap D_m)$ of hypothesis H_i $(i = 1, \ldots, k)$ given the observed data $D_1 \cap \ldots \cap D_m$ is:

$$P(H_i \mid D_1 \cap \ldots \cap D_m) = \frac{\left(\prod_{j=1}^{m} P(D_j \mid H_i)\right) P(H_i)}{P(D)}$$

where

Combining Evidence: Conditionally Independent Evidence

Definition

If we have k disjoint, exhaustive hypotheses $H_1, \ldots, H_k$ (e.g., rainy, dry) and m pieces of observed data that are conditionally independent given a hypothesis $D_1, \ldots, D_m$, then the posterior probability $P(H_i \mid D_1 \cap \ldots \cap D_m)$ of hypothesis H_i $(i = 1, \ldots, k)$ given the observed data $D_1 \cap \ldots \cap D_m$ is:

$$P(H_i \mid D_1 \cap \ldots \cap D_m) = \frac{\left(\prod_{j=1}^{m} P(D_j \mid H_i)\right) P(H_i)}{P(D)}$$

where

$$P(D) = \sum_{i=1}^{k} P(H_i) \left(\prod_{j=1}^{m} P(D_j \mid H_i)\right)$$

This Can Get You Into Trouble Sometimes

- Sally Clark was convicted in 1999 for the murder of her two infant children. Her first baby died with no evidence of foul play, so it was assumed sudden infant death syndrome (SIDS) was to blame. However, she had a second child and that baby also died. She was arrested for murder, tried, and convicted.

This Can Get You Into Trouble Sometimes

- The statistical evidence that the prosecution presented reasoned the probability of two deaths from SIDS was equal to the probability of a single death squared:

This Can Get You Into Trouble Sometimes

- The statistical evidence that the prosecution presented reasoned the probability of two deaths from SIDS was equal to the probability of a single death squared:
- $P(D_1 \cap D_2 | SIDS) = P(D_1 | SIDS) \cdot P(D_2 | SIDS)$

This Can Get You Into Trouble Sometimes

- The statistical evidence that the prosecution presented reasoned the probability of two deaths from SIDS was equal to the probability of a single death squared:
- $P(D_1 \cap D_2|SIDS) = P(D_1|SIDS) \cdot P(D_2|SIDS)$
- $P(D_1 \cap D_2|SIDS) = P(Death|SIDS)^2 =$ very small.

This Can Get You Into Trouble Sometimes

- The statistical evidence that the prosecution presented reasoned the probability of two deaths from SIDS was equal to the probability of a single death squared:
- $P(D_1 \cap D_2|SIDS) = P(D_1|SIDS) \cdot P(D_2|SIDS)$
- $P(D_1 \cap D_2|SIDS) = P(Death|SIDS)^2 =$ very small.
- However, there is evidence that if a baby dies from SIDS, the chances of it happening again are greatly increased.

This Can Get You Into Trouble Sometimes

- The statistical evidence that the prosecution presented reasoned the probability of two deaths from SIDS was equal to the probability of a single death squared:
- $P(D_1 \cap D_2|SIDS) = P(D_1|SIDS) \cdot P(D_2|SIDS)$
- $P(D_1 \cap D_2|SIDS) = P(Death|SIDS)^2 =$ very small.
- However, there is evidence that if a baby dies from SIDS, the chances of it happening again are greatly increased.
- The prosecutor also argued that since $P(D_1 \cap D_2|SIDS)$ is small, $P(SIDS|D_1 \cap D_2)$ was also small. This is a mistake because it doesn't take into account the prior probabilities of SIDS (presumably small) and murder (probably smaller!).

Classifying Spam

- Suppose you have an email and you want to know if it's spam

Classifying Spam

- Suppose you have an email and you want to know if it's spam
- In general the probability of an email being spam is 20%

Classifying Spam

- Suppose you have an email and you want to know if it's spam
- In general the probability of an email being spam is 20%
- You can compute various "features" of the email, which you can use as pieces of observed data, e.g., the presence of particular words like `viagra`, `cialis`, `cashcashcash`, ...

Classifying Spam

- Suppose you have an email and you want to know if it's spam
- In general the probability of an email being spam is 20%
- You can compute various "features" of the email, which you can use as pieces of observed data, e.g., the presence of particular words like `viagra`, `cialis`, `cashcashcash`, ...
- You have access to a lot of previously-labeled emails

Classifying Spam

- Suppose you have an email and you want to know if it's spam
- In general the probability of an email being spam is 20%
- You can compute various "features" of the email, which you can use as pieces of observed data, e.g., the presence of particular words like `viagra`, `cialis`, `cashcashcash`, ...
- You have access to a lot of previously-labeled emails
- How can you compute the probability that this email's spam?

More Formally...

- You have 2 disjoint, exhaustive hypotheses, spam and not spam, and their associated priors, $P(\text{spam})$ and $P(\text{not spam})$

More Formally...

- You have 2 disjoint, exhaustive hypotheses, spam and not spam, and their associated priors, $P(\text{spam})$ and $P(\text{not spam})$
- You have m pieces of observed data $F_1, \ldots, F_m$

More Formally...

- You have 2 disjoint, exhaustive hypotheses, spam and not spam, and their associated priors, $P(\text{spam})$ and $P(\text{not spam})$
- You have m pieces of observed data $F_1, \ldots, F_m$
- If you assume $F_1, \ldots, F_m$ are conditionally independent given the spam label, and you can compute $P(F_j \mid \text{spam})$ and $P(F_j \mid \text{not spam})$, then

More Formally...

- You have 2 disjoint, exhaustive hypotheses, spam and not spam, and their associated priors, $P(\text{spam})$ and $P(\text{not spam})$
- You have m pieces of observed data $F_1, \ldots, F_m$
- If you assume $F_1, \ldots, F_m$ are conditionally independent given the spam label, and you can compute $P(F_j \mid \text{spam})$ and $P(F_j \mid \text{not spam})$, then

$$P(\text{spam} \mid F_1 \cap \ldots \cap F_m) = \frac{\left(\prod_{j=1}^{m} P(F_j \mid \text{spam})\right) P(\text{spam})}{P(F_1 \cap \ldots \cap F_m)}$$

More Formally...

- You have 2 disjoint, exhaustive hypotheses, spam and not spam, and their associated priors, $P(\text{spam})$ and $P(\text{not spam})$
- You have m pieces of observed data $F_1, \ldots, F_m$
- If you assume $F_1, \ldots, F_m$ are conditionally independent given the spam label, and you can compute $P(F_j \mid \text{spam})$ and $P(F_j \mid \text{not spam})$, then

$$P(\text{spam} \mid F_1 \cap \ldots \cap F_m) = \frac{\left(\prod_{j=1}^{m} P(F_j \mid \text{spam})\right) P(\text{spam})}{P(F_1 \cap \ldots \cap F_m)}$$

- This equation is the basis of a naïve Bayes classifier