Machine Learning Coms-4771

Bayesian Learning
Lecture 3

January 29, 2008

(using Zoubin Ghahramani's tutorial; see the link)

Announcements

1. Homework 1 is out (due Thu, Feb 7, before class). We will go over the problems at the end of the class.
2. If you have any logistic questions about the class, email me (even though Cynthia Rudin is listed as an official instructor).

Basics

$P(X)$ – probability of X

$P(X|Y)$ – conditional probability of X given Y

$P(X, Y)$ – joint probability of X and Y

By definition:

$$P(X, Y) = P(X)P(Y|X) = P(Y)P(X|Y)$$

Bayes rule:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Marginalization:

$$P(X) = \int P(X, Y)dY$$

Bayes Rule applied to Machine Learning

The Bayesian framework:

- Assign prior belief $P(p)$ to every reasonable "guess" p
  (p is typically a process generating data; see next slide)
- Upon observing the training data S, evaluate how probable S was under each p to compute $P(S \mid p)$ (called the likelihood of S given p).
- Bayes law gives the posterior probability over "guesses" $P(p \mid S)$, which captures everything we have learned from the data:

$$P(p \mid S) = \frac{P(S \mid p)P(p)}{P(S)}$$

(decreases as $P(S)$ increases: the more probable S, independently of p, the less evidence it provides in support of p)

Prior over what?

- Generative: over models of the joint distribution $D(X, Y)$ (prior $p(Y)$ and class-conditional density $p(X \mid Y)$), typically in some parametric form; prior P is defined over the parameters θ.
- Discriminative: over models of the conditional distribution $D(Y \mid X)$ (doesn't require a prior on $D(X)$)
- Model class selection: over possible model *classes*, each parameterized by θ (model class = a distribution over distributions). Posterior for model class M:

$$P(M \mid S) = \frac{P(S \mid M)P(M)}{P(S)}$$

where marginal likelihood is given by

$$P(S \mid M) = \int P(S \mid \theta, M)P(\theta \mid M)d\theta$$

$$P(S) = \int P(S \mid M)P(M)dM$$

- Note on notation: D denotes the true (but unknown) data maker; lower-case p refers to a *model* of a data maker; capital P denotes a prior over models; In Bayesian statistics, $p(S)$ is usually denoted as $P(S \mid p) = P(S \mid \theta)$.

Predictions

- True Bayesians integrate over the posterior to make predictions:

$$P(y \mid S) = \int P(y \mid \theta, S) P(\theta \mid S) d\theta$$

- Computational shortcut (use the world with largest posterior): Maximum a posteriori (MAP) hypothesis = $\text{argmax}_\theta P(h \mid S) = \text{argmax}_\theta P(S \mid \theta) P(\theta)$.
- Further shortcut: If we assume that all θ are equally probable a priori (in some class), we only need to consider $P(S \mid \theta)$. Maximum likelihood (ML) hypothesis = $\text{argmax}_\theta P(S \mid \theta)$.
- The Dutch Book theorem: The most famous justification for the Bayesian thesis that degrees of belief should satisfy the probability calculus if we want rational behavior, i.e., should treat our beliefs the same way we treat probabilities.

  (A Dutch Book is a set of bets acceptable to the better, which are bought or sold at such prices as to always guarantee a net loss to the better, no matter what the outcome.)

Simple Example (from Mitchell's book)

- $H = \{h_1, h_2\}$, with h_1 = the patient has a certain disease and h_2 = doesn't.
- Prior: over the entire population, $P(h_1) = 0.008$ and $P(h_2) = 0.992$.
- Likelihoods (lab test result $x \in \{0, 1\}$):
  $P(x = 1 \mid h_1) = 0.98$ (true positive),
  $P(x = 0 \mid h_1) = 0.02$ (false negative),
  $P(x = 1 \mid h_0) = 0.03$ (false positive),
  $P(x = 0 \mid h_0) = 0.97$ (true negative).
- After observing positive test outcome $x = 1$, compute posteriors:
  $P(h_1 \mid x = 1) \sim P(x = 1 \mid h_1)P(h_1) = (0.98)0.008 = 0.0078$,
  $P(h_2 \mid x = 1) \sim P(x = 1 \mid h_2)P(h_2) = (0.03)0.992 = 0.0298$ (MAP)
- Normalize to 1 to get the actual probabilities (divide by $0.0078 + 0.0298$).

Properties I (informal)

- Handles the small data limit well. If our prior is correct, applying Bayes law is the optimal thing to do.
- If the observed data set S_n of size n was generated from some true data distribution given by θ^*, then under some mild conditions, and if $P(\theta^*) > 0$ (the prior probability of θ^* is non-zero), the posterior $P(\theta \mid S_n)$ will converge to the right distribution θ^* as $n \to \infty$.
- If data was generated from some distribution p^* which cannot be modelled by any θ, then the posterior will converge to some $\hat{\theta}$ minimizing the KL-divergence $\mathrm{KL}(p^*, \hat{\theta})$.

Asymptotic Properties II (informal)

Consider two Bayesians with different priors, $P_1(\theta)$ and $P_2(\theta)$, who observe the same training set S_n of size n.

Assume that both Bayesians agree on the set of possible and impossible values of θ.

$$\{\theta : P_1(\theta) > 0\} = \{\theta : P_2(\theta) > 0\}$$

Then, in the limit of $n \to \infty$, the posteriors, $P_1(\theta \mid S_n)$ and $P_2(\theta \mid S_n)$ will converge.

Bayesian Occam's Razor and Model Comparison

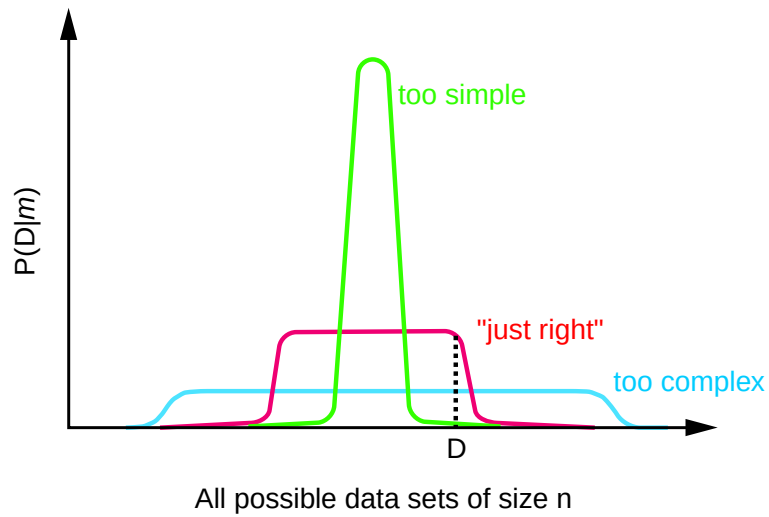
Compare model classes, e.g. m and m', using posterior probabilities given $\mathcal{D}$:

$$p(m|\mathcal{D}) = \frac{p(\mathcal{D}|m)\, p(m)}{p(\mathcal{D})}, \qquad p(\mathcal{D}|m) = \int p(\mathcal{D}|\boldsymbol{\theta}, m)\, p(\boldsymbol{\theta}|m)\, d\boldsymbol{\theta}$$

Interpretation of the Marginal Likelihood ("evidence"): The probability that *randomly selected* parameters from the prior would generate $\mathcal{D}$.

Model classes that are too simple are unlikely to generate the data set.

Model classes that are too complex can generate many possible data sets, so again, they are unlikely to generate that particular data set at random.



Each model distributes unit probability mass over all possible data sets; more probable datasets are near the center

Slide from Zoubin Ghahramani's tutorial: Notational correspondence: $\mathcal{D} = S$, $m = M$

Notes

- Computing marginal likelihoods and posteriors can be computationally difficult (if we marginalize out a number of variables, this can be a very high dimensional integral).
- There are many approximation methods (Laplace approximation, Bayesian Information Criterion, variational methods, expectation propagation, MCMC)—we won't go there (at least for now).
- Nonparametric models: Parametric models assume some finite set of parameters θ; the complexity is bounded even if the amount of data is unbounded. Nonparametric models assume an infinite dimensional θ; the complexity of the model grows with the amount of data, which makes them very flexible.

Conclusions

Bayesian methods provide a coherent framework for doing inference under uncertainty, give a language for specifying priors, and for incorporating evidence; very flexible

Limitations:

- Hard to come up with a reasonable prior, assumptions are usually wrong.
- Not very automatable, human intensive. Need to put a prior and define a set of reasonable guesses for the data before observing the data.
- Computationally difficult.

Naive Bayes Classification Algorithm

- Very simple, rarely works well
- Fat Assumption: conditioned on class, attributes are independent:

$$D(\mathbf{X} \mid Y) = \prod_i D(X_i \mid Y)$$

- Estimate marginal probabilities $D(Y = y)$ using sample frequences.
- For each label, estimate distribution of the i-th variable $D(X_i \mid Y = y)$.
- At test time, output label $\text{argmax}_y D(y \mid \mathbf{x})$ using

$$\text{argmax}_y[\log D(\mathbf{x} \mid y) + \log D(y)] = \text{argmax}_y[\log D(y) + \sum_i \log D(x_i \mid y)]$$

(where we use empirical estimates in place of probabilities)