

# **CS195-5 : Introduction to Machine Learning**

## **Lecture 10**

Greg Shakhnarovich

September 27 2006

---

# Announcements

- Problem set Nectarine out later today.
- Reminder: late hand-in of Apple till 11am on Friday!

---

# Review

- Bias-variance tradeoff: more complex  $\Rightarrow$  higher variance, lower bias.
- Representing text with binary *term occurrence* features;

$$p(\phi_j(\mathbf{x}) | y) = \theta_{jy}^{\phi_j(\mathbf{x})} (1 - \theta_{jy})^{1 - \phi_j(\mathbf{x})}.$$

- Naïve Bayes classifier assumes

$$p(\mathbf{x} | y) = p(\phi_1(\mathbf{x}), \dots, \phi_m(\mathbf{x}) | y) = \prod_{j=1}^m p(\phi_j(\mathbf{x}) | y).$$

- ML estimator  $\hat{\theta}_{jy} = \frac{1}{N} \sum_{y_i=y} \phi_j^{(i)}$  simple to compute but suffers from overfitting for small samples/rare events.

---

# Today

- Bayesian estimation: prior on parameters, Maximum A-Posteriori.
  - Conjugate priors.
  - Focus on Bernoulli / binomial case.
- Wrap up text classification with Naïve Bayes.
- Start discussion of discriminative methods.

---

# SPAM detection with Naïve Bayes

- For simplicity, we will write  $\phi_j$  instead of  $\phi_j(\mathbf{x})$ .
- SPAM corresponds to  $y = 1$ , non-SPAM to  $y = 0$ .
- NB classifier for the two-category case:

$$h_{NB}(\mathbf{x}) = 1 \Leftrightarrow \log \frac{P_1}{P_0} \prod_{j=1}^m \frac{p(\phi_j(\mathbf{x}) | y = 1)}{p(\phi_j(\mathbf{x}) | y = 0)}$$

- For a single binary feature  $\phi_j \sim \text{Bern}(\phi_j; y)$ ,

$$p(\phi_j | y = 1) = \theta_{j1}^{\phi_j} (1 - \theta_{j1})^{1 - \phi_j},$$

$$p(\phi_j | y = 0) = \theta_{j0}^{\phi_j} (1 - \theta_{j0})^{1 - \phi_j}.$$

- We need to estimate  $\theta_{j0}, \theta_{j1}$  of the Bernoulli distribution for each feature.

---

# Problems with ML estimation

- Recall the coin-tossing experiments:
  - ML is too sensitive to the data, and may violate some “reasonable” beliefs about  $\theta$ , e.g., that  $\theta = 1$  is very unlikely.
- A real problem in text classification. *Zipf's law* for English texts: the  $n$ -th most common word has relative frequency of  $1/n^a$ , with  $a \approx 1$ .
  - relative frequency means  $\#(\text{this word})/\#(\text{all words})$
- According to ML, when a word appears in a message that we have never seen in SPAM, we must decide it's legit.

---

# Problems with ML estimation

- Recall the coin-tossing experiments:
  - ML is too sensitive to the data, and may violate some “reasonable” beliefs about  $\theta$ , e.g., that  $\theta = 1$  is very unlikely.
- A real problem in text classification. *Zipf's law* for English texts: the  $n$ -th most common word has relative frequency of  $1/n^a$ , with  $a \approx 1$ .
  - relative frequency means  $\#(\text{this word})/\#(\text{all words})$
- According to ML, when a word appears in a message that we have never seen in SPAM, we must decide it's legit.
- If the same message contains a word never seen in non-SPAM, what do we do?

---

# Bayesian estimation

- The basic assumption behind the ML principle is that the unknown parameter  $\theta$  is a *fixed* quantity to be uncovered.
- An alternative, *Bayesian* view is that  $\theta$  is itself a random variable, drawn from the *parameter prior*  $p(\theta)$ .
  - The prior captures our belief about  $\theta$  *prior* to seeing any data.
- According to this view, the observed data  $X$  can be produced by *any* of the models with non-zero  $p(\theta)$ :

$$p(X) = \int_{\theta} p(X | \theta) p(\theta) d\theta.$$

- Note: we now write  $p(X | \theta)$  instead of  $p(X; \theta)$ .

---

# Frequentists versus Bayesians

- The frequentist view:  
Probability is an objective measure. It is the average frequency of an outcome if we repeat an identical experiment a large number of times.
- The Bayesian view:  
Probability is a measure of our degree of belief that a certain outcome will occur. It depends on context and may vary.
- Not to be confused with Bayes rule (used by both “camps”).

---

# Uncertainty in Bayesian estimation

- Consider a parametric model  $p(X | \theta)$  and a prior  $p(\theta)$ . Before we see  $X$ , what can we say about  $\theta$ ?

---

# Uncertainty in Bayesian estimation

- Consider a parametric model  $p(X | \theta)$  and a prior  $p(\theta)$ . Before we see  $X$ , what can we say about  $\theta$ ?
  - Only what the prior tells us.

---

# Uncertainty in Bayesian estimation

- Consider a parametric model  $p(X | \theta)$  and a prior  $p(\theta)$ . Before we see  $X$ , what can we say about  $\theta$ ?
  - Only what the prior tells us.
- After seeing the data  $X$ , our belief about  $\theta$  changes. We can describe this change using Bayes rule:

$$p(\theta | X) = \frac{1}{p(X)} p(X | \theta) p(\theta)$$

- The normalization term  $p(X) = \int_{\theta} p(X | \theta) p(\theta) d\theta$  makes sure this is still a pdf.

---

# Bayesian point estimators

$$p(\theta | X) = \frac{p(X | \theta) p(\theta)}{p(X)}$$

- We could simply stick with the *distribution over  $\theta$* .
- However, if we need to commit to a concrete estimate (a value), one reasonable choice is the *Maximum A-Posteriori* estimator:

$$\begin{aligned}\hat{\theta}_{MAP}(X) &= \operatorname{argmax}_{\theta} p(\theta | X) \\ &= \operatorname{argmax}_{\theta} p(X | \theta) p(\theta).\end{aligned}$$

- An alternative (more complicated and seldom used) approach is to estimate the expectation according to the posterior:

$$\hat{\theta}_{Exp}(X) = E_{\theta \sim p(\theta | X)} [\theta | X].$$

---

## Back to the coin tosses

$$p(\theta | X) = \frac{p(X | \theta) p(\theta)}{p(X)}, \quad \hat{\theta}_{MAP}(X) = \operatorname{argmax}_{\theta} p(X | \theta) p(\theta).$$

- We need to define a prior  $p(\theta; \alpha)$  parametrized by *hyperparameters*  $\alpha$ .
- It is convenient to use a prior such that the form of the posterior is the same as that of the prior.
  - I.e., we have a certain parametric form for our belief about  $\theta$ , and the data only changes its shape.
  - Such prior is called the *conjugate* prior for a given distribution.
- For Bernoulli (i.e. binomial likelihood) this means:

$$p(\theta; \alpha) \xrightarrow{\text{data } X} p(\theta; \alpha')$$

---

## Likelihood of $N$ Bernoulli observations

- Suppose we see a set of  $N$  observations, drawn i.i.d.  $x_i \in 0, 1 \sim \text{Bern}(x|\theta)$ .
  - We only care about the counts of heads and tails, not the order (assuming the observations are i.i.d.!).
- The likelihood of  $\theta$  given  $N_1$  successes (heads) and  $N_0$  failures (tails) in a series of independent Bernoulli trials is *binomial*:

$$\mathcal{P}(X|\theta) = \binom{N_1 + N_0}{N_1} \theta^{N_1} (1 - \theta)^{N_0}.$$

- Log-likelihood:

$$\ell(X|\theta) = \log \binom{N_1 + N_0}{N_1} + N_1 \log \theta + N_0 \log(1 - \theta).$$

---

# The Beta distribution

$$\text{Beta}(\theta; a, b) \triangleq \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1}.$$

- The *Gamma function* is a generalization of the factorial to real numbers:  $x\Gamma(x) = \Gamma(x+1)$ . The factor with  $\Gamma$ s is there for normalization.
- $\text{Beta}(\theta; a, b)$  is zero for any  $\theta \notin [0, 1]$ .
- Why is Beta conjugate for Bernoulli? Suppose  $X_N$  contains  $N_1$  ones and  $N_0$  zeros,  $N_1 + N_0 = N$ .

$$\begin{aligned} p(\theta | X_N) &\propto p(X_N | \theta) p(\theta; a, b) \\ &\propto \theta^{N_1} (1-\theta)^{N_0} \end{aligned}$$

---

# The Beta distribution

$$\text{Beta}(\theta; a, b) \triangleq \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1}.$$

- The *Gamma function* is a generalization of the factorial to real numbers:  $x\Gamma(x) = \Gamma(x+1)$ . The factor with  $\Gamma$ s is there for normalization.
- $\text{Beta}(\theta; a, b)$  is zero for any  $\theta \notin [0, 1]$ .
- Why is Beta conjugate for Bernoulli? Suppose  $X_N$  contains  $N_1$  ones and  $N_0$  zeros,  $N_1 + N_0 = N$ .

$$\begin{aligned} p(\theta | X_N) &\propto p(X_N | \theta) p(\theta; a, b) \\ &\propto \theta^{N_1} (1-\theta)^{N_0} \cdot \theta^{a-1} (1-\theta)^{b-1} \end{aligned}$$

---

# The Beta distribution

$$\text{Beta}(\theta; a, b) \triangleq \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1}.$$

- The *Gamma function* is a generalization of the factorial to real numbers:  $x\Gamma(x) = \Gamma(x+1)$ . The factor with  $\Gamma$ s is there for normalization.
- $\text{Beta}(\theta; a, b)$  is zero for any  $\theta \notin [0, 1]$ .
- Why is Beta conjugate for Bernoulli? Suppose  $X_N$  contains  $N_1$  ones and  $N_0$  zeros,  $N_1 + N_0 = N$ .

$$\begin{aligned} p(\theta | X_N) &\propto p(X_N | \theta) p(\theta; a, b) \\ &\propto \theta^{N_1} (1-\theta)^{N_0} \cdot \theta^{a-1} (1-\theta)^{b-1} \\ &= \theta^{N_1+a-1} (1-\theta)^{N_0+b-1} \end{aligned}$$

---

# The Beta distribution

$$\text{Beta}(\theta; a, b) \triangleq \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1}.$$

- The *Gamma function* is a generalization of the factorial to real numbers:  $x\Gamma(x) = \Gamma(x+1)$ . The factor with  $\Gamma$ s is there for normalization.
- $\text{Beta}(\theta; a, b)$  is zero for any  $\theta \notin [0, 1]$ .
- Why is Beta conjugate for Bernoulli? Suppose  $X_N$  contains  $N_1$  ones and  $N_0$  zeros,  $N_1 + N_0 = N$ .

$$\begin{aligned} p(\theta | X_N) &\propto p(X_N | \theta) p(\theta; a, b) \\ &\propto \theta^{N_1} (1-\theta)^{N_0} \cdot \theta^{a-1} (1-\theta)^{b-1} \\ &= \theta^{N_1+a-1} (1-\theta)^{N_0+b-1} \propto \text{Beta}(\theta; N_1+a, N_0+b) \end{aligned}$$

---

## Interpretation: pseudocounts

$$p(\theta | X_N) \propto p(\theta | N_1 + a, N_0 + b)$$

- We can think of hyperparameters  $a, b$  as *pseudocounts*:
  - Prior  $p(\theta) = \text{Beta}(\theta; a, b)$  is equivalent to having seen  $a + b$  observations,  $a$  of which were ones and  $b$  zeros, *before* we observed actual data  $X_N$ .
  - An alternative phrasing:  $a/(a + b)$  is the default value for  $\theta$ , and  $a + b$  is how strongly we believe in that value.
- The posterior  $p(\theta | X_N)$  updates that by adding the actual counts to the pseudocounts.
- As  $N \rightarrow \infty$ ,

---

## Interpretation: pseudocounts

$$p(\theta | X_N) \propto p(\theta | N_1 + a, N_0 + b)$$

- We can think of hyperparameters  $a, b$  as *pseudocounts*:
  - Prior  $p(\theta) = \text{Beta}(\theta; a, b)$  is equivalent to having seen  $a + b$  observations,  $a$  of which were ones and  $b$  zeros, *before* we observed actual data  $X_N$ .
  - An alternative phrasing:  $a/(a + b)$  is the default value for  $\theta$ , and  $a + b$  is how strongly we believe in that value.
- The posterior  $p(\theta | X_N)$  updates that by adding the actual counts to the pseudocounts.
- As  $N \rightarrow \infty$ , the effect of the prior diminishes: if  $N_1 \gg a$ ,  $N_0 \gg b$  then  $N_1 + a \approx N_1$ ,  $N_0 + b \approx N_0$ .

---

## Summary: text classification with Naïve Bayes

- Design features, e.g., based on a dictionary of  $m$  keywords.
- Compute feature representation for training set

$$\mathbf{x} \Rightarrow [\phi_1, \dots, \phi_m]^T.$$

- Estimate  $\hat{\theta}_{jy}$  for each  $j, y$  using training data.
  - ML estimation: simply count how many times  $\phi_j(\mathbf{x}_i) = 1$  for  $y_i = y$  and divide by total number of documents in class  $y$ .
  - MAP: Set a prior for each  $j = 1, \dots, m$  and  $y = 0, 1$ :  $p(\theta_{jy}) = \text{Beta}(\theta_{jy}; a_{jy}, b_{jy})$ , and find the maximum under the MAP rule.

---

# Classifying a document

- Given new document  $\mathbf{x} = [\phi_1, \dots, \phi_m]^T$ :

$$\begin{aligned}\hat{y} = 1 \Leftrightarrow & \sum_{j=1}^m \phi_j \log \theta_{j1} + \sum_{j=1}^m (1 - \phi_j) \log(1 - \theta_{j1}) \\ & - \sum_{j=1}^m \phi_j \log \theta_{j0} - \sum_{j=1}^m (1 - \phi_j) \log(1 - \theta_{j0}) \\ & + \log P_1 - \log P_0 \geq 0.\end{aligned}$$

- There are total of  $2 + 2m$  parameters to estimate in this model.