CS195-5 : Introduction to Machine Learning
Lecture 9

Greg Shakhnarovich

September 25 2006
Announcements

- Rant about exams and problem sets.
- Report how long it took you to solve PS1!
Review

• Estimation theory: estimate $\hat{\theta}$ obtained by estimator $\hat{\theta}_N$ on $X_N = x_1, \ldots, x_N$.

• The estimator is:
  – Unbiased if $\text{bias}(\hat{\theta}_N) = E[\hat{\theta}_N - \theta] = 0$;
  – Consistent if $\lim_{N \to \infty} \hat{\theta}_N = \theta$.

• Bias-variance decomposition: denote $\bar{\theta}_N = E[\hat{\theta}_N]$.

\[
E[(\hat{\theta}_N - \theta)^2] = E[(\hat{\theta}_N - \bar{\theta}_N + \bar{\theta}_N - \theta)^2]
\]
\[
= E[(\hat{\theta}_N - \bar{\theta}_N)^2] + 2(\bar{\theta}_N - \theta)E[\hat{\theta}_N - \bar{\theta}_N] + E[(\bar{\theta}_N - \theta)^2]
\]
\[
= (\bar{\theta}_N - \theta)^2 + E[(\hat{\theta}_N - \bar{\theta}_N)^2]
\]
\[
= \text{bias}^2(\hat{\theta}_N) + \text{var}(\hat{\theta}_N).
\]
Bias-variance dilemma

- **Cramer-Rao inequality**: for an unbiased estimator $\hat{\theta}_N$,

\[
\text{var}(\hat{\theta}_N) \geq \frac{1}{E\left[\left(\frac{\partial}{\partial \theta} \log p(x; \theta)\right)^2\right]}.
\]

- The **Fisher information** $E\left[\left(\frac{\partial}{\partial \theta} \log p(x; \theta)\right)^2\right]$ is related to the shape of $p(x; \theta)$. Intuitively, it measures the amount of information data $X$ provides about parameter $\theta$. 
Bias-variance in regression

- The true model: \( y = F(x) + \nu \), zero-mean additive noise \( \nu \).
  - \( F \) not necessarily \( \in \mathcal{F} \)

- We estimate \( \hat{w} \) from \( X_N \) and approximate \( F(x) \) with \( f(x; \hat{w}) \).

- Denote:
  - The average of \( f(x; \hat{w}) \) over training sets \( X_N \):
    \[
    \bar{f}(x) = E_{X_N} [f(x; \hat{w})]
    \]
  - The best estimate with a function \( \in \mathcal{F} \):
    \[
    f^*(x) = f(x; \arg\min_w E_{p(x,y)} [(y - f(x; w))^2])
    \]
  - An estimated \( f \) on a particular \( X_N \):
    \[
    \hat{f}(x) = f(x; \hat{w})
    \]
Bias-variance in regression

- Focus on a single $x_0$:

$$EX_N [(y_0 - \hat{f}(x_0))^2] = EX_N [(y_0 - \bar{f}(x_0))^2] + EX_N [(\hat{f}(x_0) - \bar{f}(x_0))^2].$$

- The second term is the variance of $f$.

- The first term can be further decomposed:

$$EX_N [(y_0 - \bar{f}(x_0))^2] = EX_N [(y_0 - F(x_0))^2] + EX_N [(F(x_0) - \bar{f}(x_0))^2].$$

  - The irreducible error $E [(y_0 - F(x_0))^2]$, due to noise variance.
  - The bias$^2$ term $E [(F(x_0) - \bar{f}(x_0))^2]$, due to difference between $f$ and $F$.

Need to integrate all of this over $x_0, y_0$ to get the expected bias and variance.
Model complexity and bias-variance

- Model complexity can be measured by the number of independent parameters to be fit (“degrees of freedom”).

- For instance, Gaussian with full covariance $\Rightarrow d(d+1)/2$ parameters; diagonal covariance $\Rightarrow d$ parameters, spherical covariance $\Rightarrow 1$ parameter.

- As an informal rule, the bias-variance tradeoff is as follows:
  - Complex model $\Rightarrow$ sensitive to data $\Rightarrow$ much affected by changes in $X_N \Rightarrow$ high variance, low bias.
  - Simple model $\Rightarrow$ more rigid $\Rightarrow$ does not change as much with changes in $X_N \Rightarrow$ low variance, high bias.

- One of the most important goals in learning: finding a model that is just right in the bias-variance tradeoff.
Naïve Bayes classifier

- Also called “Idiot’s Bayes”.
- Suppose $\mathbf{x}$ is represented by $m$ features $\phi_1(\mathbf{x}), \ldots, \phi_m(\mathbf{x})$.
  - The simplest case: $\phi_j(\mathbf{x}) = x_j$
- NB assumes that the features are *independent* given the class:

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

- Under this assumption, the Bayes classifier is

$$h^*(\mathbf{x}) = \text{sign} \left[ \sum_{j=1}^{m} \log \frac{p(\phi_j(\mathbf{x}) | +1)}{p(\phi_j(\mathbf{x}) | -1)} + \log P_{+1} - \log P_{-1} \right].$$
Naïve Bayes for Gaussian model

\[ p(x | c) = p(\phi_1(x), \ldots, \phi_m(x) | c) = \prod_{j=1}^{m} p(\phi_j(x) | c). \]

- \( \phi_j(x) = x_j \); NB assumption of independence is equivalent to

\[ \Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \ldots & 0 \\ 0 & \sigma_2^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & \ldots & \sigma_d^2 \end{bmatrix} \]

- Need to estimate the \( d \) marginal 1D Gaussian densities (one for each component of \( x \)).
Example: generative models for documents

• A common task: given an e-mail message, classify it as SPAM or “ham” (a legitimate e-mail).

• Define a set of keywords $W_1, \ldots, W_m$.

$$\phi_j(x) = \begin{cases} 1 \quad \text{document } x \text{ includes } W_j, \\
0 \quad \text{otherwise.} \end{cases}$$

• A document $x$ (of arbitrary length!) is now represented as a vector in $\{0, 1\}^m$:
$$\Phi(x) = [\phi_1(x), \ldots, \phi_m(x)]^T.$$  

• A natural distribution for $\phi_j(x)$ is Bernoulli.
Discrete probability distributions

- Often the observations are discrete by nature:
  - Text documents;
  - Genetic code;
  - Binary images (silhouettes).

- Makes some of the math simpler:
  - Probability mass function instead of probability density.
  - Sums instead of integrals:
    \[
    \sum_{v \in X} p(x = v) = 1.
    \]
    \[
    E[f(x)] = \sum_{v \in X} f(v)p(x = v).
    \]
Bernoulli random variables

- A single binary variable, i.e. \( X = \{0, 1\} \), parametrized by \( \theta \):
  \[
p(x = 1; \theta) = \theta.
  \]
e.g., a single flip of a coin with \( \text{Prob(\text{heads})} = \theta \).

- An alternative form: \( p(x|\theta) = \theta^x (1 - \theta)^{1-x} \).
  \[
  E[x] = \theta, \quad \text{var}(x) = \theta (1 - \theta).
  \]

- ML estimate of \( \theta \) from a set \( N \) observations \( x_1, \ldots, x_N \):
  \[
  \hat{\theta}_{ML} = \frac{1}{N} \sum_{i=1}^{N} x_i.
  \]
Extending Bernoulli to more than two values

• Suppose $x$ assumes values in $\{1, \ldots, K\}$, with $p(x = k) = \theta_k$.

• A trick: represent $x$ with a $K$-bit vector $x$. E.g.,

$$K = 6, \ x = 3 \quad \Rightarrow \quad x = [0, 0, 1, 0, 0, 0]^T.$$  

• Then, denote $\theta = [\theta_1, \ldots, \theta_K]^T$, we get

$$p(x; \theta) = \prod_{k=1}^{K} \theta_k^{x_k}.$$  

E.g., $p([0, 0, 1, 0, 0, 0]^T; \theta) = \theta_1^0 \cdot \theta_2^0 \cdot \theta_3^1 \cdot \theta_4^0 \cdot \theta_5^0 \cdot \theta_6^0 = \theta_3$.  

Multinomial distribution

• Suppose we have $N$ values drawn from the $K$-valued distribution parametrized by $\theta = [\theta_1, \ldots, \theta_K]^T$. Let

$$N_k = \text{number of times the value } k \text{ appears.}$$

Of course, $\sum_k N_k = N$.

• The distribution of the $K$-dimensional vector $\mathbf{n} = [N_1, \ldots, N_K]$ is called Multinomial:

$$p(\mathbf{n}; \theta) = \binom{N}{N_1, \ldots, N_K} \theta_1^{N_1} \cdots \theta_K^{N_K}.$$
Application: SPAM detection

- Given an e-mail message need to classify it as SPAM \((y = 1)\) or “ham” \((y = 0)\), based on the content.

- An important problem! \(P_1\) pretty high...

- Typical binary features:
  - keywords;
  - HTML tags and patterns;
  - SCREAMING LINES (ALL CAPS);
  - number of recipients above certain threshold;
  - Comes from “blacklisted” relay...
SPAM detection with Naïve Bayes

• For simplicity, we will write $\phi_j$ instead of $\phi_j(x)$.

• For a single binary feature $\phi_j$,

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

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

• We need to estimate $\theta_{j0}, \theta_{j1}$ for each feature.

• ML estimate of a Bernoulli variable:
  – Suppose we have observed $k$ heads and $N - k$ tails.
  – ML estimate of $\theta$ is $k/N$.  

Problems with ML estimation

- Suppose we have tossed a coin three times; denote H = 1, T = 0.
- We want to estimate the coin's $\theta = p(1)$.
- The resulting sequence: $x_1 = 1$,
Problems with ML estimation

• Suppose we have tossed a coin three times; denote $H = 1$, $T = 0$.

• We want to estimate the coin's $\theta = p(1)$.

• The resulting sequence: $x_1 = 1$, $x_2 = 1$, 
Problems with ML estimation

• Suppose we have tossed a coin three times; denote $H=1$, $T=0$.

• We want to estimate the coin's $\theta = p(1)$.

• The resulting sequence: $x_1 = 1$, $x_2 = 1$, $x_3 = 1$. 
Problems with ML estimation

- Suppose we have tossed a coin three times; denote $H=1$, $T=0$.
- We want to estimate the coin’s $\theta = p(1)$.
- The resulting sequence: $x_1 = 1$, $x_2 = 1$, $x_3 = 1$.
- ML estimate:
Problems with ML estimation

• Suppose we have tossed a coin three times; denote H = 1, T = 0.

• We want to estimate the coin’s $\theta = p(1)$.

• The resulting sequence: $x_1 = 1, x_2 = 1, x_3 = 1$.

• ML estimate: $\hat{\theta} = 1$ (coin heavily bent).
Problems with ML estimation

- Suppose we have tossed a coin three times; denote H = 1, T = 0.

- We want to estimate the coin’s $\theta = p(1)$.

- The resulting sequence: $x_1 = 1, x_2 = 1, x_3 = 1$.

- ML estimate: $\hat{\theta} = 1$ (coin heavily bent).

- However, if the coin is fair ($\theta = 0.5$) we should expect this outcome in $1/8$ of such experiments...
  - The ML estimate is unbiased, but has pretty high variance.
Problems with ML estimation

- Suppose we have tossed a coin three times; denote $H = 1$, $T = 0$.

- We want to estimate the coin's $\theta = p(1)$.

- The resulting sequence: $x_1 = 1$, $x_2 = 1$, $x_3 = 1$.

- ML estimate: $\hat{\theta} = 1$ (coin heavily bent).

- However, if the coin is fair ($\theta = 0.5$) we should expect this outcome in $1/8$ of such experiments...
  - The ML estimate is unbiased, but has pretty high variance.

- We need to formalize our intuition that a perfectly bent coin is not very likely...
Bayesian estimation

• The basic assumption behind the ML principle is that the unknown parameters $\theta$ 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 or 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”).**
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:

\[ \hat{\theta}_{MAP}(X) = \operatorname{argmax}_\theta p(\theta | X) \]
\[ = \operatorname{argmax}_\theta p(X | \theta) p(\theta). \]

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

\[ \hat{\theta}_{Exp}(X) = \mathbb{E}_{\theta \sim p(\theta | X)} [\theta | X]. \]
Back to the coin tosses

\[
\hat{\theta}_{MAP}(X) = \arg\max_{\theta} p(X | \theta) p(\theta).
\]

- What prior should we use?

- It is convenient to use a prior such that the form of the posterior is the same as that of the prior.
  - Depends on the form of the likelihood;
  - Such prior is called conjugate prior for a given likelihood.

- For Bernoulli likelihood, the Beta prior is conjugate.
Next time

Conjugate priors.
Finish NB example for documents.
Priors for Gaussian models in 1D and $d$-D.
Discriminative learning.