

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 $X_N = \mathbf{x}_1, \dots, \mathbf{x}_N$.
- The estimator is:
 - Unbiased if $\text{bias}(\hat{\theta}_N) = E[\hat{\theta}_N - \theta] = 0$;
 - Consistent if $\lim_{N \rightarrow \infty} \hat{\theta}_N = \theta$.
- Bias-variance decomposition: denote $\bar{\theta}_N = E[\hat{\theta}_N]$.

$$\begin{aligned} E[(\hat{\theta}_N - \theta)^2] &= E[(\hat{\theta}_N - \bar{\theta}_N + \bar{\theta}_N - \theta)^2] \\ &= E[(\hat{\theta}_N - \bar{\theta}_N)^2] + 2(\bar{\theta}_N - \theta)E[\hat{\theta}_N - \bar{\theta}_N] + E[(\bar{\theta}_N - \theta)^2] \\ &= (\bar{\theta}_N - \theta)^2 + E[(\hat{\theta}_N - \bar{\theta}_N)^2] \\ &= \text{bias}^2(\hat{\theta}_N) + \text{var}(\hat{\theta}_N). \end{aligned}$$

Bias-variance dilemma

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

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

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

Bias-variance in regression

- The true model: $y = F(\mathbf{x}) + \nu$, zero-mean additive noise ν .
 - F not necessarily $\in \mathcal{F}$
- We estimate $\hat{\mathbf{w}}$ from X_N and approximate $F(\mathbf{x})$ with $f(\mathbf{x}; \hat{\mathbf{w}})$.
- Denote:
 - The average of $f(\mathbf{x}; \hat{\mathbf{w}})$ over training sets X_N :

$$\bar{f}(\mathbf{x}) = E_{X_N} [f(\mathbf{x}; \hat{\mathbf{w}})]$$

- The best estimate with a function $\in \mathcal{F}$:

$$f^*(\mathbf{x}) = f(\mathbf{x}; \underset{\mathbf{w}}{\operatorname{argmin}} E_{p(\mathbf{x}, y)} [(y - f(\mathbf{x}; \mathbf{w}))^2])$$

- An estimated f on a particular X_N :

$$\hat{f}(\mathbf{x}) = f(\mathbf{x}; \hat{\mathbf{w}})$$

Bias-variance in regression

- Focus on a single \mathbf{x}_0 :

$$E_{X_N} \left[(y_0 - \hat{f}(\mathbf{x}_0))^2 \right] = E_{X_N} \left[(y_0 - \bar{f}(\mathbf{x}_0))^2 \right] + E_{X_N} \left[(\hat{f}(\mathbf{x}_0) - \bar{f}(\mathbf{x}_0))^2 \right].$$

- The second term is the *variance* of f .
- The first term can be further decomposed:

$$E_{X_N} \left[(y_0 - \bar{f}(\mathbf{x}_0))^2 \right] = E_{X_N} \left[(y_0 - F(\mathbf{x}_0))^2 \right] + E_{X_N} \left[(F(\mathbf{x}_0) - \bar{f}(\mathbf{x}_0))^2 \right] :$$

- The *irreducible error* $E \left[(y_0 - F(\mathbf{x}_0))^2 \right]$, due to noise variance.
- The *bias*² term $E \left[(F(\mathbf{x}_0) - \bar{f}(\mathbf{x}_0))^2 \right]$, due to difference between f and F .

Need to integrate all of this over \mathbf{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 $\phi_1(\mathbf{x}), \dots, \phi_m(\mathbf{x})$.
 - The simplest case: $\phi(j)(\mathbf{x}) = x_j$
- NB assumes that the features are *independent* given the class:

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

- 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

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

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

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & \dots & \sigma_d^2 \end{bmatrix}$$

- Need to estimate the d marginal 1D Gaussian densities (one for each component of \mathbf{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, \dots, W_m .

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

- A document \mathbf{x} (of arbitrary length!) is now represented as a vector in $\{0, 1\}^m$:
 $\Phi(\mathbf{x}) = [\phi_1(\mathbf{x}), \dots, \phi_m(\mathbf{x})]^T$.
- A natural distribution for $\phi_j(\mathbf{x})$ 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:

$$\sum_{v \in \mathcal{X}} p(x = v) = 1.$$

$$E[f(x)] = \sum_{v \in \mathcal{X}} f(v)p(x = v).$$

Bernoulli random variables

- A single binary variable, i.e. $\mathcal{X} = \{0, 1\}$, parametrized by θ :

$$p(x = 1; \theta) = \theta.$$

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 θ from a set N observations x_1, \dots, 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 $\{1, \dots, K\}$, with $p(x = k) = \theta_k$.
- A trick: represent x with a K -bit vector \mathbf{x} . E.g.,

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

- Then, denote $\theta = [\theta_1, \dots, \theta_K]^T$, we get

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

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

Multinomial distribution

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

N_k = number of times the value k appears.

Of course, $\sum_k N_k = N$.

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

$$p(\mathbf{n}; \theta) = \binom{N}{N_1, \dots, N_K} \theta_1^{N_1} \dots \theta_K^{N_K}.$$

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 ϕ_j instead of $\phi_j(\mathbf{x})$.
- For a single binary feature ϕ_j ,

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

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

- We need to estimate θ_{j0}, θ_{j1} for each feature.
- ML estimate of a Bernoulli variable:
 - Suppose we have observed k heads and $N - k$ tails.
 - ML estimate of θ 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 θ is a *fixed* quantity to be uncovered.
- An alternative, *Bayesian* view is that θ is itself a random variable, drawn from the *parameter prior* $p(\theta)$.
 - The prior captures or belief about θ *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 θ ?

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 θ ?
 - 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 θ ?
 - Only what the prior tells us.
- After seeing the data X , our belief about θ 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 θ* .
- 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

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

- 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.