Estimating Probabilities

Léon Bottou

NEC Labs America

COS 424 – 2/25/2010

Today's Agenda

Goals	Classification, clustering, regression, other.
Representation	Parametric vs. kernels vs. nonparametric Probabilistic vs. nonprobabilistic Linear vs. nonlinear Deep vs. shallow
Capacity Control	Explicit: architecture, feature selection Explicit: regularization, priors Implicit: approximate optimization Implicit: bayesian averaging, ensembles
Operational Considerations	Loss functions Budget constraints Online vs. offline
Computational Considerations	Exact algorithms for small datasets. Stochastic algorithms for big datasets. Parallel algorithms.

Introduction

Direct Method

(1) Minimize a loss that is direcly related to our goal.

Probabilistic Method

(1) Estimate probabilities.

(2) Use estimated probabilities to implement our goal(s).

Drawbacks

– Estimating probabilities may be more difficult than solving our goal.

– Additional steps bring new opportunities for error.

Benefits

– Improved ability to *reason* about the data.

– Multiple goals.

Summary

1. Estimating probabilities and densities.

2. Maximum Likelihood

3. Comparing estimators

4. Classical approach
   – Unbiased estimators

5. Bayesian approach
   – An alternate view on probabilities
   – Priors and posteriors
   – Averaging

6. Putting them together!

Estimating a probability

Estimate $p = \mathbb{P}_X\{X \in A\}$ given a sample $x_1, \dots, x_n$.

Represent the possible samples

– *Independent and identically distributed* random variables

$$\mathbb{P}\{X_1, \dots, X_n\} = \mathbb{P}_X\{X_1\}\,\mathbb{P}_X\{X_2\}\dots\mathbb{P}_X\{X_n\}$$

Law of large numbers, etc.

– For instance with the CLT: $\bar{X} \;=\; \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{X_i \in A\} \;\sim\; \mathcal{N}\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$

therefore $\mathbb{P}\left\{ \left| \bar{X} - p \right| \leq 2\sqrt{\frac{p(1-p)}{n}} \right\} \approx 95\%$ etc.

Notes:

– The 95% mean 95% of the possible samples.

– Estimating a single probability works nicely.

Estimating a cumulative distribution

Estimate $F(x) = \mathbb{P}_X\{X \leq x\}$ given a sample $x_1, \dots, x_n$.

Represent the possible samples

– *Independent and identically distributed* random variables

$$\mathbb{P}\{X_1, \dots, X_n\} = \mathbb{P}_X\{X_1\}\,\mathbb{P}_X\{X_2\}\dots\mathbb{P}_X\{X_n\}$$

Glivenko-Cantelli

– Let $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X \leq x)$.

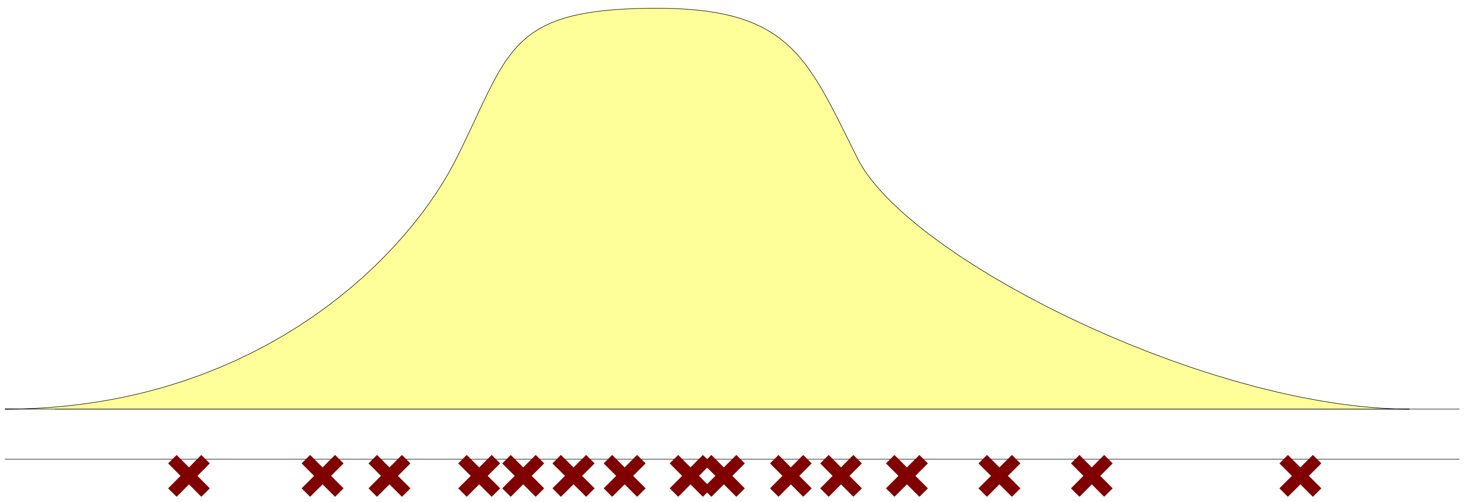
– Then $\mathbb{P}\left\{ \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| > \epsilon \right\} \leq Ce^{-2n\epsilon^2}$

Notes:

– This is not an obvious result.

– Estimating a cumulative distribution works nicely.

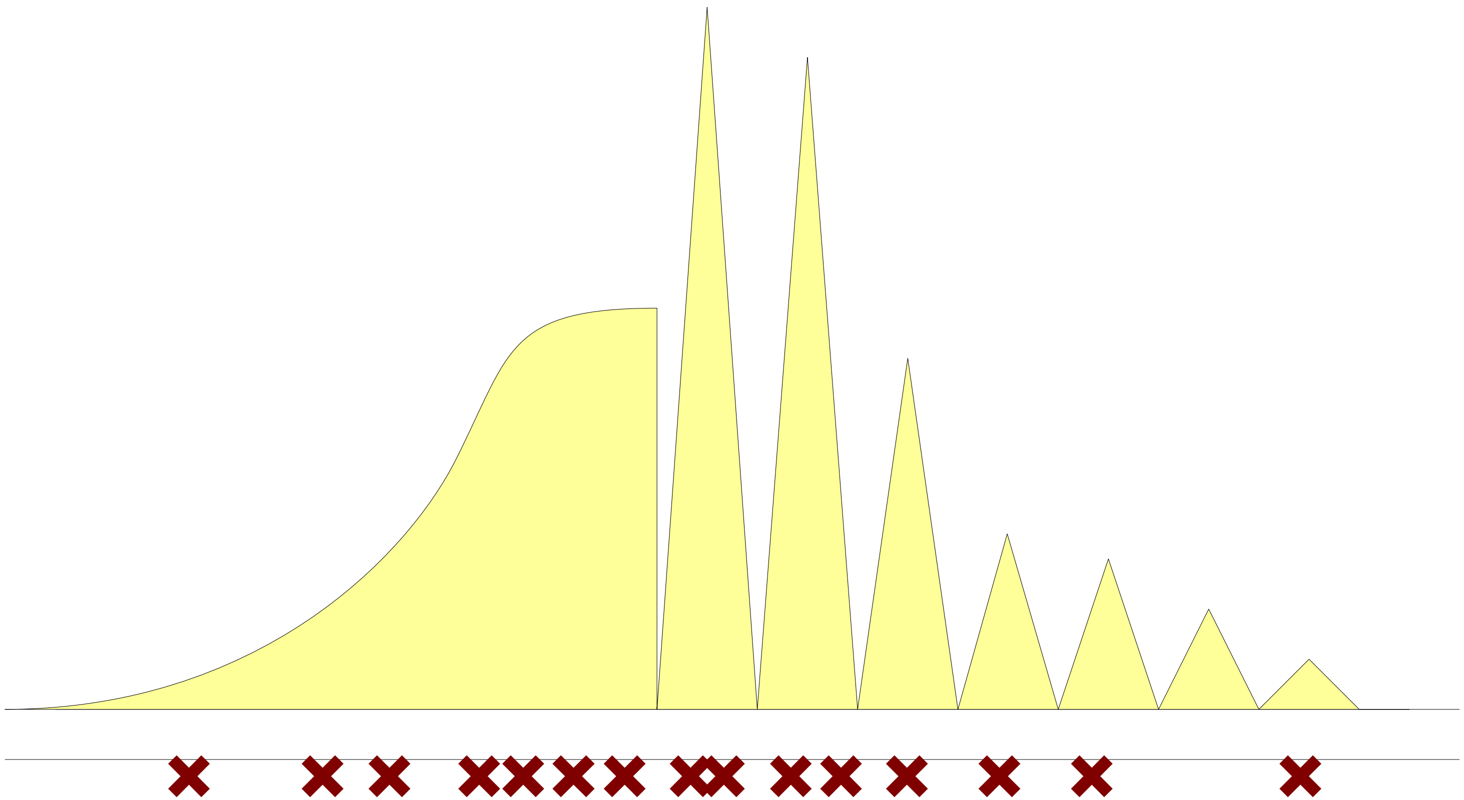
Estimating a density



Notes:

– The density is the derivative of the cumulative.

Estimating a density



Notes:

– The density is the derivative of the cumulative.

– Estimating a density is nearly impossible.

A convenient shortcut

Assume we know the distribution up to a few parameters θ.

	Discrete	Continuous
Parametric form	$\mathbb{P}\{X = x\} = f_{\theta^*}(x)$	$p(x) = f_{\theta^*}(x)$
Normalization	$\sum_x f_\theta(x) = 1$	$\int p(x)\,dx = 1$

Likelihood

– $L(\theta; x_1 \dots x_n) \ \stackrel{\Delta}{=}\ \prod_{i=1}^{n} f_\theta(x_i)$

i.e. the probability of $x_1 \dots x_n$ if f_θ was the real distribution.

Maximum Likelihood Estimator (MLE)

– $\hat{\theta} \ \stackrel{\Delta}{=}\ \arg\max_\theta L(\theta; x_1 \dots x_n) \ = \ \arg\max_\theta \sum_{i=1}^{n} \log f_\theta(x_i)$

MLE for the Bernoulli distribution

– X takes value 1 with probability p and value 0 with probability $1-p$.
– Estimate p from a sample $x_1, \dots, x_n$ with n_1 ones and n_0 zeroes.

Likelihood

– $L(p) = p^{n_1}\,(1-p)^{n_0}$
– $\log L(p) = n_1 \log(p) + n_0 \log(1-p)$.

Maximum Likelihood

– $\dfrac{d \log L}{dp} = \dfrac{n_1}{p} - \dfrac{n_0}{1-p} = 0$ gives $\boldsymbol{\hat{p} = \dfrac{n_1}{n_1 + n_0}}$

MLE for the Normal distribution

– Assume $X \sim \mathcal{N}(\mu, \sigma)$.
– Estimate μ and σ from a sample $x_1, \dots, x_n$.

Likelihood

– Let $\gamma = 1/\sigma$.

– $L(\mu,\sigma) = \prod_{i=1}^{n} \frac{\gamma}{\sqrt{2\pi}} e^{-\frac{1}{2}\gamma^2 (x_i-\mu)^2}$

– $\log L(\mu,\sigma) = n \log \gamma - \frac{\gamma^2}{2} \sum_{i=1}^{n} (x_i - \mu)^2$

Maximum Likelihood

– $\frac{d \log L}{d\mu} = \sum_{i=1}^{n} (x_i - \mu) = 0$ gives $\boldsymbol{\mu = \frac{1}{n} \sum_{i=1}^{n} x_i}$

– $\frac{d \log L}{d\gamma} = \frac{n}{\gamma} - \gamma \sum_{i=1}^{n} (x_i - \mu)^2 = 0$ gives $\boldsymbol{\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2}$

Why does MLE work

Kullback-Leibler divergence

– Between discrete distributions: $D(P\|Q) = \sum P(x) \log(P(x)/Q(x))$

– Between probability densities: $D(p\|q) = \int p(x) \log(p(x)/q(x))\, dx$

The KL-divergence measures how p differs from q

– Since $\log(x) \leq x - 1$, $D(p\|q) \geq \int p(x) \left[\frac{q(x)}{p(x)} - 1\right] dx = \int p(x)\, dx - \int q(x)\, dx = 1 - 1 = 0$.

– $D(p\|q) = 0$ if and only if $p = q$.

MLE and KL-divergence

– Observe $\frac{1}{n} \log L(\theta) \stackrel{n\to\infty}{\longrightarrow} \int p(x) \log f_\theta(x)\, dx = \text{Constant} - D(p\|f_\theta)$

– Therefore MLE approaches $\arg\min_\theta D(p\|f_\theta)$.

– Same when $p(x)$ does not have the assumed parametric form.

MLE for classification

Generative

– Let $p_\theta(x,y)$ estimate $\mathbb{P}\{X \mid Y=y\}$.

– Required normalization: $\forall y,\theta,\ \int p_\theta(x,y)dx = 1$.

– Maximum likelihood: $\max \frac{1}{n}\sum_{i=1}^{n} \log p_\theta(x_i,y_i)$.

Discriminative

– Let $p_\theta(x,y)$ estimate $\mathbb{P}\{Y=y \mid X\}$.

– Required normalization: $\forall x,\theta,\ \sum_y p_\theta(x,y) = 1$.

– Maximum likelihood: $\max \frac{1}{n}\sum_{i=1}^{n} \log p_\theta(x_i,y_i)$.

Only the normalization differs!

MLE for binary classification

Let $p_\theta(x)$ estimate $\mathbb{P}\{Y = +1 \mid X\}$.

The log-likelihood is $\log L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \begin{cases} \log(p_\theta(x_i)) & \text{if } y_i = +1 \\ \log(1 - p_\theta(x_i)) & \text{if } y_i = -1 \end{cases}$

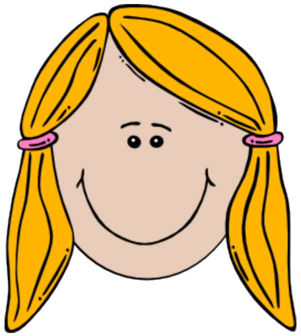
Observe $\log L(\theta) \;=\; -\frac{1}{n} \sum_{i=1}^{n} \log\left(1 + e^{-y_i z_\theta(x)}\right)$ with $z_\theta(x) = \log \frac{p_\theta(x)}{1 - p_\theta(x)}$.

We recover a classifier with the log loss!

Conversely, when using the log-loss to train a classifier $f(x)$,
the quantities $\frac{e^{f(x)}}{1+e^{f(x)}}$ and $\frac{1}{1+e^{f(x)}}$ approximate $\mathbb{P}\{Y = \pm 1 \mid X\}$.

Comparing estimators

Estimate $\mathbb{E}[X]$ given a sample $x_1, \ldots, x_n$.



$$\mathbb{E}[X] \approx \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Jane believes in hard labor.

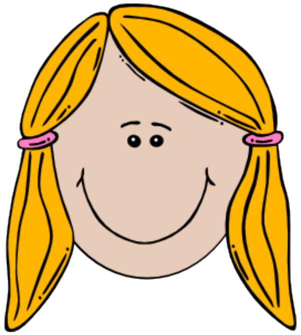
$$\mathbb{E}[X] \approx 3 \;!$$

Joe does not.

Is Jane's answer always better than Joe's ?

Comparing estimators

Estimate $\mathbb{E}[X]$ given a sample $x_1, \ldots, x_n$.



$$\mathbb{E}[X] \approx \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Jane believes in hard labor.

$$\mathbb{E}[X] \approx 3 \ !$$

Joe does not.

Is Jane's answer always better than Joe's ?

– There are probability distributions $\mathbb{P}\{X\}$ whose expectation is 3.
– For these, Joe is exactly right (because he is lucky.)
– And Jane is likely to answer 2.98 or 3.01. . .

Can we at least say that Jane is right more often?

– Only if we can say which distributions are more likely to occur. . .

A philosophical debate

Bayesian: *Let us just fix a probability distribution on the possible probability distributions of X. We'll call that the prior.*

Classical: *There is no such thing. You can only count occurrences of X. You cannot count probability distributions.*

Bayesian: *Does it matter? Let's just say that the prior represents my a priori beliefs about the problem.*

Classical: *Where did you get these beliefs from? Are you telling me that the probability distribution of X is partly known beforehand? You are cheating.*

Bayesian: *Well, my beliefs could be right or wrong. The important thing is to be consistent.*

Classical: *You might be consistently wrong.*

Bayesian: *Maybe I'll change my mind when I see enough data.*

Classical approach: no lucky Joes.

We want to estimate $\mu \in \mathbb{R}$ that depends on the distribution of X.
We do that with an estimator $\hat{\mu}(x_1, x_2, \ldots, x_n)$.

Unbiased estimator

$\mathbb{E}\,[\,\hat{\mu}(X_1, \ldots, X_n)\,] = \mu$ regardless of the distribution of X.

Examples

– $\bar{x} = \frac{1}{n}\sum x_i$ is an unbiased estimator of $\mu = \mathbb{E}\,[X]$.

– $\bar{v} = \frac{1}{n-1}\sum (x_i - \bar{x})^2$ is an unbiased estimator of $\sigma^2 = \text{Var}(X)$.

because
$$\begin{aligned}
\mathbb{E}\left[\sum (X_i - \bar{X})^2\right] &= \mathbb{E}\left[\sum \left((X_i - \mu) - (\bar{X} - \mu)\right)^2\right] \\
&= \mathbb{E}\left[\sum (X_i = \mu)^2 \;-\; 2(\bar{X} - \mu)\sum (X_i - \mu) \;+\; n(\bar{X} - \mu)^2\right] \\
&= n\sigma^2 - n\mathbb{E}\left[(\bar{X} - \mu)^2\right] \;=\; n\sigma^2 - n\mathbb{E}\left[\left(\sum \frac{X_i - \mu}{n}\right)^2\right] \\
&= n\sigma^2 - \frac{1}{n}\mathbb{E}\left[\sum \text{Var}(X_i - \mu)\right] \;=\; (n-1)\sigma^2
\end{aligned}$$

Classical approach: no lucky Joes.

Best unbiased estimator

– There are optimal unbiased estimators that are uniformly better than all other unbiased estimators.

– Deriving the best unbiased estimator is often very difficult.

– MLE is only asymptotically unbiased and asymptotically efficient.

Is unbiasedness a good idea?

– What if we actually have a priori information ?

– A priori information can take subtle forms.

Stein's paradox (1961)

– The batting averages y_i of different players are independent.

– Best unbiased estimators: $\hat{y}_i = \#\text{hits}_i / \#\text{bats}_i$

– Let $\bar{y}$ be a grand average and c an appropriate shrinking factor.

– Biased estimators: $\hat{x}_i = \bar{y} + c(\hat{y}_i - \bar{y})$.

– On average over all players, $\hat{x}$ is uniformly better than $\hat{y}$.

Bayesian approach: no unknown probability.

Probabilities in classical statistics

– Probabilities $\mathbb{P}\{\dots\}$ represent the unknown.

  "Unknown probability distribution $\mathbb{P}\{X\}$"

  "Discover something about $\mathbb{P}\{X\}$ using a sample"

  "Regardless of the actual distribution. . . "

– Likelihoods $p_\theta(x)$ behave like probabilities but represent models.

Probabilities in Bayesian statistics

– Probabilities $\mathbb{P}\{\dots\}$ represent our beliefs.

– There are no unknown probabilities: we know what our beliefs are!

– The classical likelihood $p_\theta(x)$ is similar to the Bayesian $\mathbb{P}\{X \mid \theta\}$.

– We can have beliefs $\mathbb{P}\{\theta\}$ about θ.

Both are unfortunately represented with the same letter $\mathbb{P}$.

Learning with Bayes rule.

Prior information

– The model: $\mathbb{P}\{X \mid \theta\}$.

– The prior distribution: $\mathbb{P}\{\theta\}$.

Posterior distribution

– We observe some data $D = \{x_1, x_2, \ldots, x_n\}$.

– Applying Bayes rule:

$$\begin{aligned}
\mathbb{P}\{\theta \mid D\} &= \mathbb{P}\{D \mid \theta\}\,\mathbb{P}\{\theta\} \,/\, \mathbb{P}\{D\} \\
&\propto \mathbb{P}\{D \mid \theta\}\,\mathbb{P}\{\theta\} \\
&\propto \mathbb{P}\{X_1 \mid \theta\}\,\mathbb{P}\{X_2 \mid \theta\} \ldots \mathbb{P}\{X_n \mid \theta\}\,\mathbb{P}\{\theta\}
\end{aligned}$$

Averaging

– Then $\mathbb{P}\{X \mid D\} = \displaystyle\int \mathbb{P}\{X \mid \theta\}\,\mathbb{P}\{\theta \mid D\}\,d\theta$

Bayes for the Bernoulli distribution

Prior information

– The model: $\mathbb{P}\{X = x \mid \theta\} = \begin{cases} \theta & \text{if } x = 1 \\ 1-\theta & \text{if } x = 0 \end{cases}$

– The prior distribution: $\mathbb{P}\{\theta\} \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$ for $\alpha, \beta > 0$.

Posterior distribution

– We observe $D = \{x_1, x_2, \ldots, x_n\}$ with n_1 ones and n_0 zeroes.

– Applying Bayes rule:

$$\begin{aligned} \mathbb{P}\{\theta \mid D\} &\propto \mathbb{P}\{X_1 \mid \theta\}\,\mathbb{P}\{X_2 \mid \theta\} \ldots \mathbb{P}\{X_n \mid \theta\}\,\mathbb{P}\{\theta\} \\ &\propto \theta^{n_1+\alpha-1}(1-\theta)^{n_0+\beta-1} \end{aligned}$$

Bayes for the Bernoulli distribution (2)

Useful special functions

– Gamma function: $\Gamma(x) = (x-1)\Gamma(x-1)$.

– Beta function: $B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$

Averaging

– $\mathbb{P}\{X=+1 \mid D\} \propto B(n_1+\alpha+1, n_0+\beta) = \frac{n_1+\alpha}{n_1+\alpha+n_0+\beta}B(n_1+\alpha, n_0+\beta)$

– $\mathbb{P}\{X=-1 \mid D\} \propto B(n_1+\alpha, n_0+\beta+1) = \frac{n_0+\beta}{n_1+\alpha+n_0+\beta}B(n_1+\alpha, n_0+\beta)$

Conclusion: $\mathbb{P}\{X=1 \mid D\} = \dfrac{n_1+\alpha}{n_1+\alpha+n_0+\beta}$

– Same as MLE but initialize counts to $\alpha, \beta > 0$.

– Large α, β bias the probability towards $\alpha/(\alpha+\beta)$.

– The influence of the prior vanishes when n increases.

– Prior is a capacity control device.

Remarks about Bayesian statistics

Relation to MLE

– MLE always has an uniform prior

– MLE takes $\theta = \arg\max \mathbb{P}\{\theta \mid D\}$ instead of averaging.

Computation of the Bayesian averages

– Analytical: *Conjugate priors* make the derivations less hairy.

– Approximate: *Laplace approximation* summarizes the posterior.

– Numerical: *Markov-Chain Monte Carlo* and variants.

Putting things together

Lets use different letters:

– $\mathbb{Q}$ is the classical (unknown) probability,

– $\mathbb{P}$ is the Bayesian probability (or the classical likelihood.)

The MLE question: $\mathbb{P}\{X \mid \theta = \arg\max \mathbb{P}\{\theta \mid D\}\} \to \mathbb{Q}\{X\}$?

i.e. Is MLE consistent?

– With discrete probabilities: **yes**.

– With continuous probabilities: **often**.

The Bayesian question: $\mathbb{P}\{X \mid D\} \to \mathbb{Q}\{X\}$?

i.e. Do the priors vanish when n increases?

– With discrete probabilities: **yes**.

– With continuous probabilities: **more often than MLE**.