Notes of Machine Learning Foundation

Xiangli Chen

Computer Science Department
University of Illinois at Chicago

July 27, 2017

Outline

Set

Definition

A σ-field (σ-algebra) $\mathcal{B}$ is a non-empty class of subsets of Ω closed under countable union, countable intersection and complements.

A minimal set of postulates for $\mathcal{B}$ to be a σ-field is

1. $\Omega \in \mathcal{B}$
2. $B \in \mathcal{B}$ implies $B^c \in \mathcal{B}$
3. $B_i \in \mathcal{B}, i \geq 1$ implies $\cup_{i=1}^{\infty} B_i \in \mathcal{B}$

Let $\mathcal{C}$ be a collection of subsets of Ω. The σ-field generated by $\mathcal{C}$, $\sigma(\mathcal{C})$ (minimal σ-filed over $\mathcal{C}$), is a σ-field satisfying $\sigma(\mathcal{C}) \supset \mathcal{C}$ and $\mathcal{B}' \supset \sigma(\mathcal{C})$ if $\mathcal{B}'$ is some other σ-field containing $\mathcal{C}$.
Suppose $\Omega = \mathbb{R}$ and let $\mathcal{C} = \{(a, b], -\infty \leq a \leq b < \infty\}$, define $\mathcal{B}(\mathbb{R}) := \sigma(\mathcal{C})$ and call $\mathcal{B}(\mathbb{R})$ the Borel subsets of $\mathbb{R}$.

Probability

Random experiment: can be repeated under the same condition, and each experiment terminates with an outcome

Sample space Ω: the collection of every possible outcome.

Definition

A probability space is a triple $(\Omega, \mathcal{B}, p)$ where

- Ω is the sample space corresponding to outcomes of some experiment.
- $\mathcal{B}$ is the σ-field of subsets (events) (may be not all) of Ω.
- p is a probability measure $p : \mathcal{B} \to [0, 1]$, a function such that
  1. $p(A) \geq 0$ for all $A \in \mathcal{B}$
  2. p is σ-additive: If $\{A_n, n \geq 1\}$ are events in $\mathcal{B}$ that are disjoint, then
  $$p(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} p(A_n).$$
  3. $p(\Omega) = 1$

An Example

This is an example where σ-field not equals the power set: $\mathcal{B} \neq \mathcal{P}(\Omega)$

Example

Let $\Omega = \{0, 1\}^{\mathbb{N}}$ be the set of sequences with values only 0 and 1. Let $\mathcal{B} = \{\emptyset, B_0, B_1, \Omega\}$ where $B_i = \{\omega \in \Omega : \omega_1 = i\}$ (the first value of the sequence ω is i).
Then $\mathcal{B}$ is the σ-field of subsets of Ω. If we define $p(B_0) = p(B_1) = 1/2$, then $(\Omega, \mathcal{B}, p)$ defnies a probability space.

Random Variable

Measurable space: a pair $(\Omega, \mathcal{B})$ consists of a set and a σ-field of subsets. If $(\Omega, \mathcal{B})$ and $(\Omega', \mathcal{B}')$ are two measurable spaces, then a map

$$X : \Omega \to \Omega'$$

is called measurable if

$$X^{-1}(\mathcal{B}') \subset \mathcal{B}.$$

X is also called a random element of Ω'.

Definition

X is called a random variable when $(\Omega', \mathcal{B}') = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Let $(\Omega, \mathcal{B}, p)$ be a probability space, define the function $p \circ X^{-1}$ on $\mathcal{B}'$ by

$$p \circ X^{-1}(A') = p(X^{-1}(A')).$$

$p \circ X^{-1}$ is a probability on $(\Omega', \mathcal{B}')$ called the induced probability or the distribution of X.

Distribution and Density Functions

Definition

The cumulative distribution function (cdf) $F_X(x)$ of a random variable X is

$$F_X(x) = p_X((-\infty, x]) = p(\{c \in \mathcal{C} : X(c) \leq x\}).$$

We say a random variable X is continuous if its cdf $F_X(x)$ is a continuous function for all $x \in R$.

Definition

Most continuous random variables are absolutely continuous; i.e.,

$$F_X(x) = \int_{-\infty}^{x} f_X(t) d_t$$

The function $f_X(t)$ is called a probability density function (pdf) of X. If $f_X(x)$ is also continuous, then

$$\frac{d}{dx} F_X(x) = f_X(x).$$

Transformations

Theorem

Let X be a continuous random variable with pdf $f_X(x)$ and support S_X. Let $Y = g(X)$, where $g(x)$ is a one-to-one differentiable function on S_X. We note that $x = g^{-1}(y)$ and let $dx/dy = d(g^{-1}(y))/dy$. Then

$$f_Y(y) = f_X(g^{-1}(y))|\frac{dx}{dy}|, \text{for } y \in S_Y,$$

where the support of Y is $S_Y = \{y = g(x) : x \in S_X\}$.

Transformation proof

Proof.

Since $g(x)$ is one-to-one and continuous, it is either strictly monotonically increasing or decreasing. Assume it is strictly monotonically increasing,

$$F_Y(y) = P[Y \leq y] = P[g(X) \leq y] = P[X \leq g^{-1}(y)] = F_X(g^{-1}(y)).$$

$$f_Y(y) = \frac{d}{dy}F_Y(y) = f_X(g^{-1}(y))\frac{dx}{dy}, \quad (\frac{dx}{dy} > 0)$$

Suppose it is strictly monotonically decreasing,

$$F_Y(y) = 1 - F_X(g^{-1}(y))$$

$$f_Y(y) = \frac{d}{dy}F_Y(y) = f_X(g^{-1}(y))(-\frac{dx}{dy}) \quad (\frac{dx}{dy} < 0),$$

□

Important Inequalities

Theorem

Markov's Inequality

Let $u(X)$ be a nonnegative function. If $\mathbb{E}[u(X)]$ exists, then for any $c > 0$,

$$p(u(X) \geq c) \leq \mathbb{E}[u(X)]/c.$$

Chebyshev's Inequality

Assume there is a finite variance σ^2 (implies $\mu = \mathbb{E}$ exists), then for envery $k > 0$,

$$p(|X - \mu| \geq k\sigma) \leq 1/k^2.$$

Jensen's Inequality

If ϕ is convex on an open interval I and the support of X is contained in I and $\mathbb{E}[X] < \infty$, then

$$\phi([\mathbb{E}(X)]) \leq \mathbb{E}[\phi(X)].$$

If ϕ is strictly convex, then the inequality is strict unless X is a constant.

Outline

1. Probability
   - Theory
   - Some Special Distributions
   - Asymptotic Theory

2. Statistical Inference
   - Confidence Interval
   - Frequentist Statistics
     - Maximum Likelihood Estimation (MLE)
     - Expectation and Maximization Algorithm (EM)
   - Bayesian Statistics

Bernoulli, Binomial and Multinomial Distributions

Bernoulli (e.g. flip coin) $\quad X \sim \text{ber}(\theta)$

$$p(x) = \theta^x (1-\theta)^{1-x} \quad x = 0, 1 \quad \mu = \theta \quad \sigma^2 = \theta(1-\theta)$$

Binomial (e.g. number of successes) $\quad X \sim bin(n, \theta)$

$$p(x) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \quad x = 0, 1, 2, \ldots, n$$

$$\mu = n\theta \qquad \sigma^2 = n\theta(1-\theta)$$

Let X_i's iid and $X_i \sim \text{bin}(n_i, \theta)$, then $Y = \sum_{i=1}^m X_i \sim \text{bin}(\sum_{i=1}^m n_i, \theta)$.
Multinomial (e.g. tossing a K-side die) $\quad X_j \sim mul(n, \theta_j)$

$$p(x_1, \ldots, x_K) = \frac{n!}{\prod_{j=1}^K x_j!} \prod_{j=1}^K \theta_j^{x_j} \quad \sum_{j=1}^K x_j = n \;\; x_j \in \mathbb{N} \quad \sum_{j=1}^K \theta_j = 1$$

Poisson Distribution

Poisson (e.g. number of alpha paricles, defects, automobile accidents)

$$X \sim \text{poi}(\lambda) \quad \lambda > 0$$
$$p(x) = e^{-\lambda}\frac{\lambda^x}{x!} \quad x \in \mathbb{N}$$
$$\mu = \sigma^2 = \lambda$$

Note that

$$1 + \lambda + \frac{\lambda^2}{2!} + \cdots = \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{\lambda}$$

Let X_i's iid and $X_i \sim \text{poi}(\lambda_i)$, then $Y = \sum_{i=1}^n X_i \sim \text{poi}(\sum_{i=1}^n \lambda_i)$.

Gamma Distribution

Gamma function

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} d_x = \begin{cases} (\alpha-1)\Gamma(\alpha-1) & \alpha > 1 \\ 1 & \alpha = 1 \end{cases}$$

If $\alpha > 1$ and $\alpha \in \mathbb{N}$,

$$\Gamma(\alpha) = (\alpha-1)(\alpha-2)\cdots 1\Gamma(1) = (\alpha-1)!.$$

This suggests $0! = 1$.
Gamma (e.g. time needed to obtain α changes modeled with poisson)

$$X \sim \Gamma(\alpha, \beta) \quad \alpha > 0 \quad \beta > 0$$

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta} \quad 0 < x < \infty \quad \mu = \alpha\beta \quad \sigma^2 = \alpha\beta^2$$

Let X_i's iid and $X_i \sim \Gamma(\alpha_i, \beta)$, then $Y = \sum_{i=1}^n X_i \sim \Gamma(\sum_{i=1}^n \alpha_i, \beta)$.

Exponential, Laplace, Chi-square distributions

Exponential $X \sim e(\lambda) = \Gamma(1, 1/\lambda) \quad \lambda > 0$

$$f(x) = \lambda e^{-\lambda x} \quad 0 < x < \infty \quad \mu = 1/\lambda \quad \sigma^2 = 1/\lambda^2$$

Laplace (double sided exponential)

$$X \sim \text{lap}(\mu, b) \quad -\infty < \mu < \infty \quad b > 0$$
$$f(x) = 1/(2b)e^{-\frac{|x-\mu|}{b}} \quad -\infty < x < \infty \quad \mu = \mu \quad \sigma^2 = 2b^2$$

Chi-squared

$$X \sim \mathcal{X}^2(r) = \Gamma(r/2, 2) \quad r > 0$$
$$f(x) = \frac{1}{\Gamma(r/2)2^{r/2}} x^{r/2-1} e^{-x/2} \quad 0 < x < \infty \quad \mu = r \quad \sigma^2 = 2r$$

Let X_i's iid, $X_i \sim \mathcal{X}^2(r_i)$, then $Y = \sum_{i=1}^n X_i \sim \mathcal{X}^2(\sum_{i=1}^n r_i)$.

Beta, Dirichlet distributions

Beta

Let $X_1 \sim \Gamma(\alpha, 1)$, $X_2 \sim \Gamma(\beta, 1)$ and $X_1 \perp X_2$, define $X = X_1/(X_1 + X_2)$,

$$X \sim \beta(\alpha, \beta) \quad \alpha > 0, \quad \beta > 0$$

$$f(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha - 1}(1 - x)^{\beta - 1} \quad 0 < x < 1$$

$$\mu = \frac{\alpha}{\alpha + \beta} \quad \sigma^2 = \frac{\alpha\beta}{(\alpha + \beta + 1)(\alpha + \beta)^2}$$

Beta function

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}$$

Dirichlet (built from $\Gamma(\alpha_j, 1)$)

$$f(x_1, \ldots, x_K) = \frac{1}{B(\alpha_1, \ldots, \alpha_K)} \prod_{j=1}^{K} x_j^{\alpha_j - 1} \quad 0 \le x_j \le 1 \quad \sum_{j=1}^{K} x_j = 1$$

Normal Distribution

If $X \sim N(\mu, \sigma^2), \sigma^2 > 0$, then $V = (X - \mu)^2/\sigma^2 \sim \mathcal{X}^2(1)$.
To be continued

Multivariate Normal Distribution

To be continued

t and F Distributions

t

Let $W \sim N(0,1), V \sim \mathcal{X}^2(r)$ and $W \perp V$,

$$T = W/(\sqrt{V/r}) \sim t(r) \quad r > 0$$

$$f(x) = \frac{\Gamma((r+1)/2)}{\sqrt{\pi r}\Gamma(r/2)(1+x^2/r)^{(r+2)/2}}, \quad -\infty < x < \infty$$

$$\mu = 0 \quad r > 1; \quad \sigma^2 = r/(r-2) \quad r > 2$$

F

Let $U \sim \mathcal{X}^2(r_1), V \sim \mathcal{X}^2(r_2)$ and $U \perp V$,

$$F = (U/r_1)/(V/r_2) \sim F(r_1, r_2) \quad r_1 > 0 \quad r_2 > 0$$

$$f(x) = \frac{\Gamma((r_1+r_2)/2)(r_1/r_2)^{r_1/2} x^{r_1/2-1}}{\Gamma(r_1/2)\Gamma(r_2/2)(1+r_1 x/r_2)^{(r_1+r_2)/2}} \quad x > 0$$

$$\mu = r_2/(r_2-2) \quad r_2 > 2; \quad \sigma^2 = 2\left(\frac{r_2}{r_2-2}\right)^2 \frac{r_1+r_2-2}{r_1(r_2-4)} \quad r_2 > 4$$

Student's Theorem

Let $X_1, \cdots, X_n$ iid and $X_i \sim N(\mu, \sigma^2)$, define

$$\text{Sample mean: } \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\text{Sample variance: } S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

- $\overline{X} \sim N(\mu, \sigma^2/n)$
- $\overline{X} \perp S^2$
- $(n-1)S^2/\sigma^2 \sim \mathcal{X}^2(n-1)$
- $T = \frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$

Mixture Distributions

To be continued

Outline

Basic Concepts

- sample: $X_1, \ldots, X_n$ have the same distribution
- random sample: $X_1, \ldots, X_n$ iid
- statistics: a function of a sample $T = T(X_1, \ldots, X_n)$

Converge in probability

Definition

Let $\{X_n\}$ be a sequence of random variables and let X be a random variable. We say that X_n converges in probability to X if, for all $\epsilon > 0$,

$$\lim_{n\to\infty} p((x_n - x) \geq \epsilon) = 0$$

or equivalently,

$$\lim_{n\to\infty} p((x_n - x) < \epsilon) = 1.$$

If so, we write

$$X_n \xrightarrow{p} X.$$

Converge of the real sequence $a_n \to a$ is equivalent to $a_n \xrightarrow{p} a$.

- Suppose $X_n \xrightarrow{p} X$, $Y_n \xrightarrow{p} Y$, then $X_n + Y_n \xrightarrow{p} X + Y, X_n Y_n \xrightarrow{p} XY$.
- Suppose $X_n \xrightarrow{p} X$, then $aX_n \xrightarrow{p} aX$.
- Suppose $X_n \xrightarrow{p} X$ and g is continuous, then $g(X_n) \xrightarrow{p} g(X)$.

Law of Large Numbers

Definition

Weak law of large numbers
$\{X_n\}$ iid, mean $\mu < \infty$, variance $\sigma^2 < \infty$, let

$$\overline{X}_n = n^{-1} \sum_{i=1}^{n} X_i$$

Then

$$\overline{X}_n \xrightarrow{p} \mu$$

Strong law of large numbers only requires $\mu < \infty$.

Consistency

Definition

Consistency:
Let X be a random variable with cdf $F(x, \theta)$, $\theta \in \Theta$. Let $X_1, \cdots, X_n$ be a sample from the distribution of X and let T_n denote a statistic. We say T_n is a consistent estimator of θ if

$$T_n \xrightarrow{p} \theta.$$

Example

Consistent estimator

- Sample mean (iid) $\overline{X}_n \xrightarrow{p} \mu \quad \mu < \infty$
- Sample variance (iid) $S_n^2 \xrightarrow{p} \sigma^2, S_n \xrightarrow{p} \sigma \quad \mathrm{var}(S^2) < \infty$

Converge in Distribution

Definition

Let $\{X_n\}$ be a sequence of random variables and let X be a random variable. Let F_{X_n} and F_X be, respectively, the cdfs of X_n and X. Let $C(F_X)$ denote the set of all points where F_X is continuous. We say that X_n converges in distribution to X if

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x), \text{ for all } x \in C(F_X).$$

We denote this convergence by If so, we write

$$X_n \xrightarrow{D} X.$$

Example

- Let $Y_n \sim \text{bin}(n, p)$, we know $\mu = np$, then $Y_n \xrightarrow{D} \text{poi}(\mu)$.
- Let $Z_n \sim \mathcal{X}^2(n)$, then $Y_n = (Z_n - n)\sqrt{(}2n) \xrightarrow{D} N(0, 1)$.

Some properties

- If $X_n \xrightarrow{p} X$, then $X_n \xrightarrow{D} X$.
- If $X_n \xrightarrow{D} b$ (a constant), then $X_n \xrightarrow{p} b$.
- If $X_n \xrightarrow{p} X$, $Y_n \xrightarrow{p} 0$, then $X_n + Y_n \xrightarrow{D} X$.
- If $X_n \xrightarrow{p} X, g$ is continuous, then $g(X_n) \xrightarrow{D} g(X)$

Theorem

Slutsky's Theorem
If $X_n \xrightarrow{D} X, A_n \xrightarrow{p} a, B_n \xrightarrow{p} b$ (a, b constant), then $A_n + B_n X_n \xrightarrow{D} a + bX$

Bounded in Probability

Definition

Bounded in probability (Stochastically bounded):
We say that the sequence of random variables $\{X_n\}$ is bounded in probability if, for all $\epsilon > 0$, there exists a constant $C > 0$ and an integer N_ϵ such that for all $n \geq N_\epsilon$

$$p(|X_n| \leq C) \geq 1 - \epsilon.$$

Some properties

- If $X_n \xrightarrow{D} X$, then $\{X_n\}$ is bounded in probability.
- $\{X_n\}$ is bounded in probability and $Y_n \xrightarrow{p} 0$, then $X_n Y_n \xrightarrow{p} 0$

Converge Rate

Consider random sequences $\{X_n\}_{n=1}^{\infty}$ and $\{Y_n\}_{n=1}^{\infty}$.

The o_p notation.

$X_n = o_p(Y_n)$ if and only if

$$\frac{X_n}{Y_n} \xrightarrow{p} 0.$$

The O_p notation.

$X_n = O_p(Y_n)$ (X_n is of order no larger than Y_n) if and only if

$$\frac{X_n}{Y_n} = O_p(1).$$

If $X_n = O_p(1)$, we say that X_n is bounded in probability.

If $\{Y_n\}$ is bounded in probability and $X_n = o_p(Y_n)$, then $X_n \xrightarrow{p} 0$

Δ-Method

Theorem

Δ *method*
Let $\{X_n\}$ *be a sequence of random variables such that*

$$\sqrt{n}(X_n - \theta) \xrightarrow{D} N(0, \sigma^2).$$

Suppose the function $g(x)$ *is differentiable at* θ *and* $g'(\theta) \neq 0$. *Then*

$$\sqrt{n}(g(X_n) - g(\theta)) \xrightarrow{D} N(0, \sigma^2(g'(\theta))^2).$$

Central Limit Theory

Theorem

Central limit theory:
Let $X_1, \ldots, X_n$ *iid from a distribution has mean* μ *and variance* $\sigma^2 > 0$. *Then*

$$Y_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{D} N(0,1)$$

Remark

Note $\mathbb{E}[\overline{X}_n] = \mu$ and $\mathrm{Var}(\overline{X}_n) = \sigma^2/n$, CLT shows $\overline{X}_n \xrightarrow{D} N(\mu, \sigma^2/n)$.

Large sample inference for μ - $X_1, \ldots, X_n$ iid, μ and σ^2 are unknown,

$$\frac{\overline{X}_n - \mu}{S/\sqrt{n}} \xrightarrow{D} N(0,1)$$

Note that

$$\frac{\overline{X}_n - \mu}{S/\sqrt{n}} = \left(\frac{\sigma}{S}\right) \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \text{ and } s \xrightarrow{p} \sigma.$$

CLT applications

Normal approximation to the binomial distribution
Let $X_1, \ldots, X_n$ iid, $X_i \sim \text{ber}(p)$, we know $Y_n = X_1 + \ldots + X_n \sim \text{bin}(n, p)$

$$\frac{\overline{X}_n - p}{\sqrt{p(1-p)/n}} \xrightarrow{D} N(0, 1)$$

Large sample inference for properties
Let $X_1, \ldots, X_n$ iid, $X_i \sim \text{ber}(p)$, let $\hat{p} = \bar{(}X)$

$$\frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}} \xrightarrow{D} N(0, 1)$$

Note that $\hat{p} = (\overline{X}) \xrightarrow{p} p$
Large sample inference for $\mathcal{X}^2$-test
Let $Y_n \sim \text{bin}(n, p)$, we know $(Y_n - np)/\sqrt{np(1-p)} \xrightarrow{D} N(0, 1)$. Then

$$((Y_n - np)/\sqrt{np(1-p)})^2 \xrightarrow{D} \mathcal{X}^2(1)$$

Outline

Confidence Interval

Definition

Confidence Interval Let $X_1, \ldots, X_n$ be a sample on $X \sim f(x; \theta), \theta \in \Omega$.
Let $L = L(X_1, \ldots, X_n)$ and $U = U(X_1, \ldots, X_n)$ be two statistics and $0 < \alpha < 1$.
The interval (L, U) is a $(1-\alpha)100\%$ confidence interval for θ if

$$1 - \alpha = p_\theta(\theta \in (L, U))$$

Confidence interval for μ-under normality
$X_1, \ldots, X_n$ iid $X_i \sim N(\mu, \sigma^2)$, we know $T = (\overline{X} - \mu)(s/\sqrt(n)) \sim t(n-1)$.

$$1 - \alpha = p(\overline{x} - t_{\alpha/2,n-1} S/\sqrt{n} < \mu < \overline{x} + t_{\alpha/2,n-1} S/\sqrt{n})$$

Some Applications

Large sample confidence interval for μ
$X_1, \ldots, X_n$ iid, X_i has mean μ and variance σ^2, we know
$\overline{X}_n - \mu/(S/\sqrt{n}) \xrightarrow{D} Z = N(0,1)$

$$1 - \alpha \approx p\left(\overline{x} - z_{\alpha/2} S/\sqrt{n} < \mu < \overline{x} + z_{\alpha/2} S/\sqrt{n}\right)$$

Large sample confidence interval for p
$X_1, \ldots, X_n$ iid, $X_i \sim \text{ber}(p)$, let $\hat{p} = \overline{X}$. Note $\text{Var}(\hat{p}) = p(1-p)/n$, by CLT $(\hat{p} - p)\sqrt{p(1-p)/n} \xrightarrow{D} Z = N(0,1)$ and we know $\hat{p} \xrightarrow{D} p$.

$$1 - \alpha \approx p\left(\hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} < \mu < \hat{p} + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}\right)$$

Outline

Maximum Likelihood Estimation (MLE)

MLE: Choose θ that maximizes the probability of observed data $\mathcal{D}$

$$\hat{\theta}_{MLE} = \operatorname*{argmax}_{\theta} p(\mathcal{D}|\theta)$$

In general, we assume $\mathcal{D} : x_1, \cdots, x_n$ independent draw and identically distributed (iid) of $p(x|\theta)$.

$$\hat{\theta}_{MLE} = \operatorname*{argmax}_{\theta} \prod_{i=1}^{n} p(x_i|\theta)$$

Let $J(\theta) = \prod_{i=1}^{n} p(x_i|\theta)$, often,

$$\partial J(\theta)/\partial\theta|_{\theta=\hat{\theta}_{MLE}} = 0$$

In pratice, for computation benefit, we consider the log form, $L(\theta)$

MLE-Examples

To be continued

Expectation and Maximization Algorithm (EM)

EM: a general method of finding the maximum likelihood estimate.
Motivation: when MLE of a problem is analytically intractable and gradient learning method is slow.
Approach: computing the MLE of an incomplete-data problem by formulating an associated complete-data problem.
Remark: If local maximums of likelihood exist, EM algorithm may monotonically converge to a local maximum.
Application: Gaussian mixture model, hidden markov model

Formulation

Data

- observed $x = (x_1, \ldots, x_{n_1})$
- unobserved $z = (z_1, \ldots, z_{n_2})$

Likelihood

- observed $f(x|\theta)$ complete $f(x, z|\theta)$
- conditional $f(z|x, \theta) = f(x, z|\theta)/f(x|\theta)$

E-M algorithm: imporve $f(x|\theta)$ using $f(x, z|\theta)$.

$$
\begin{aligned}
\log f(x|\theta) &= \int (\log f(x|\theta)) f(z|x, \theta') dz \\
&= \int (\log f(x, z|\theta) - \log f(z|x, \theta)) f(z|x, \theta') dz \\
&= \mathbb{E}_{z|x,\theta'}(\log f(x, Z|\theta)|x, \theta') - \mathbb{E}_{z|x,\theta'}(\log f(Z|x, \theta)|x, \theta')
\end{aligned}
$$

EM Algorithm

Initialize the estimate $\hat{\theta}^0$ of θ. Let $\hat{\theta}^m$ be the estimate on the m-th step. To compute the estimate $\hat{\theta}^{m+1}$ on the $m+1$-th step, do

1. Expectation step (E step)

$$Q(\theta|\hat{\theta}^m, x) = \mathbb{E}_{z|x,\hat{\theta}^m}(\log f(x, Z|\theta)|x, \hat{\theta}^m)$$

2. Maximization step (M step)

$$\hat{\theta}^{m+1} = \arg\max Q(\theta|\hat{\theta}^m, x)$$

EM algorithm monotonically improves the likelihood $f(x|\theta)$ (or unchanged).

Theorem

The sequence of estimate $\hat{\theta}^m$, defined by EM algorithm, satisfies

$$\log f(x|\hat{\theta}^{m+1}) \geq \log f(x|\hat{\theta}^m)$$

EM Property

Proof.

$$
\begin{aligned}
&\log f(x|\hat\theta^{m+1}) - \log f(x|\hat\theta^{m}) \\
&\quad = \int \log f(x|\hat\theta^{m+1}) f(z|x,\hat\theta^{m})dz - \int \log f(x|\hat\theta^{m}) f(z|x,\hat\theta^{m})dz \\
&\quad = \int (\log f(x,z|\hat\theta^{m+1}) - \log f(z|x,\hat\theta^{m+1})) f(z|x,\hat\theta^{m})dz - \\
&\qquad \int (\log f(x,z|\hat\theta^{m}) - \log f(z|x,\hat\theta^{m})) f(z|x,\hat\theta^{m})dz \\
&\quad = Q(\hat\theta^{m+1}|\hat\theta^{m},x) - \mathbb{E}_{z|x,\hat\theta^m}(\log f(Z|x,\hat\theta^{m+1})|x,\hat\theta^{m}) - \\
&\qquad (Q(\hat\theta^{m}|\hat\theta^{m},x) - \mathbb{E}_{z|x,\hat\theta^m}(\log f(Z|x,\hat\theta^{m})|x,\hat\theta^{m})) \\
&\quad = \underbrace{Q(\hat\theta^{m+1}|\hat\theta^{m},x) - Q(\hat\theta^{m}|\hat\theta^{m},x)}_{\hat\theta^{m+1}=\arg\max Q(\theta|\hat\theta^{m},x)} + \underbrace{\mathbb{E}_{z|x,\hat\theta^m}\left(\log \frac{f(Z|x,\hat\theta^{m})}{f(Z|x,\hat\theta^{m+1})} \Big| x,\hat\theta^{m}\right)}_{\text{KL divergence}\geq 0} \geq 0
\end{aligned}
$$

EM in General

Factorize the log likelihood as

$$\begin{aligned}\log f(x|\theta) &= \int (\log f(x|\theta))\, f(z)dz \\ &= \int (\log f(x,z|\theta) - \log f(z|x,\theta))\, f(z)dz \\ &= L(f(z),\theta) + KL(f(z)||f(z|x,\theta))\end{aligned}$$

where

$$L(f(z),\theta) = \int \log\left(\frac{f(x,z|\theta)}{f(z)}\right) f(z)dz$$

$$KL(f(z)||f(z|x,\theta)) = \int \log\left(\frac{f(z)}{f(z|x,\theta)}\right) f(z)dz$$

Lower Bound

$L(f(z), \theta)$ is a lower bound on $\log f(x|\theta)$ (KL divergence is nonegative)

$$L(f(z), \theta) \leq \log f(x|\theta)$$

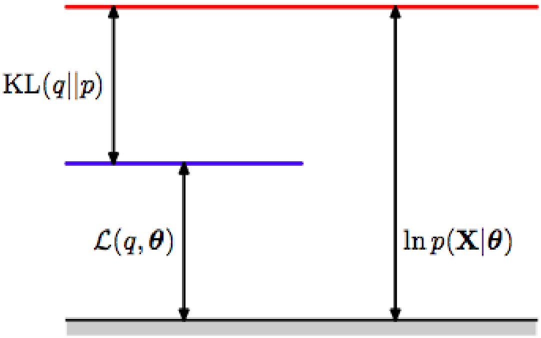

Figure 1: lower bound

Interprete E-step

E-step is equivalent to (suppose the current estimate is $\hat{\theta}^m$)

$$\begin{aligned}
\hat{f}^{m+1}(z) &= \arg\max_{f(z)} L(f(z), \hat{\theta}^m) \\
&= \arg\min_{f(z)} \log f(x|\hat{\theta}^m) - L(f(z), \hat{\theta}^m \\
&= \arg\min_{f(z)} KL(f(z)||f(z|x, \hat{\theta}^m)) = f(z|x, \hat{\theta}^m)
\end{aligned}$$

In this case, $KL(.) = 0$, i.e. $L(\hat{f}^{m+1}(z)) = \log f(x|\hat{\theta}^m)$.

Interprete E-step

E-step causes $L() = \log f()$

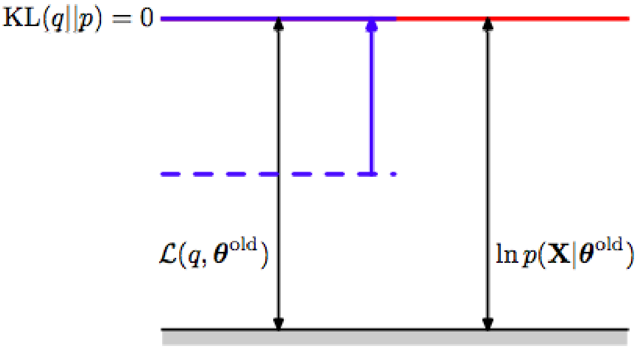

Figure 2: E-step

Interprete M-step

M-step is equivalent to

$$\begin{aligned}
\hat{\theta}^{m+1} &= \arg\max_{\theta} L(f(z|x, th\hat{e}ta^{m}), \theta) \\
&= \arg\max_{\theta} \int (\log(f(x, z|\theta)))f(z|x, \hat{\theta}^{m})dz + \text{const} \\
&= \arg\max_{\theta} Q(\theta, \hat{\theta}^{m}) + \text{const}
\end{aligned}$$

The log likelihood becomes

$$\log f(x|\hat{\theta}^{m+1}) = L(f(z|x, \hat{\theta}^{m}), \hat{\theta}^{m+1}) + KL(f(z|x, \hat{\theta}^{m})||f(z|x, \hat{\theta}^{m+1}))$$

Interprete M-step

The M-step causes both $L(.)$ and $\log f(.)$ increase, and KL divergence becomes nonzero (unless converges to maximum).

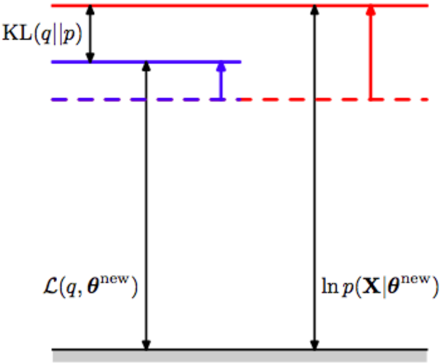

Figure 3: M-step

Operations of the EM Algorithm

EM algorithm

- E-step: $L(.) = \log f(.)$ and $KL(.) = 0$
- M-step: $\log f(.) \uparrow = L(.) \uparrow + KL(.) (\geq 0)$

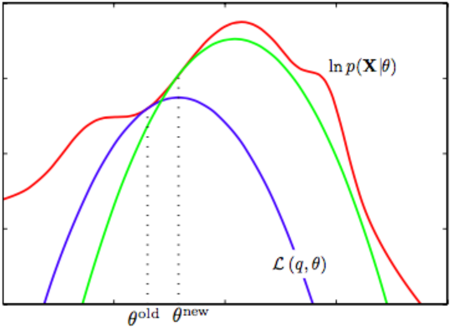

Figure 4: EM-algorithm

Converge Property

EM-algorithm overview

$$\hat{\theta} = \arg\max_{\theta} \max_{f(z)} L(f(z), \theta) = \arg\max_{\theta} f(x|\theta)$$

- Any algorithm that converges to the global maximum of $L(., \theta)$ that is also a global maximum of $\log f(x|\theta)$.
- Provided $f(x, z|\theta)$ is continuous on θ, local maximum of $L(., \theta)$ is also a local maximum of $\log f(x|\theta)$.

EM Algorithm for MAP

Posterior distribution

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{f(x)}$$

$$\begin{aligned}\log f(\theta|x) &= \log f(x|\theta) + \log f(\theta) - f(x) \\ &= L(f(z), \theta) + KL(f(z)||f(z|x, \theta)) + \log f(\theta) - f(x)\end{aligned}$$

EM algorithm

- E-step: the same
- M-step: consider additional $\log f(\theta)$

Outline

Maximum a Posteriori Estimation (MAP)

A Bayesian approach

$$p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta)$$

where prior $p(\theta)$ represents expert knowledge.
Conjugate priors: $p(\theta)$ and $p(\theta|\mathcal{D})$ have the same form.
Maximum a posteriori (MAP) estimation: choose value that is most probable given observed data and prior belief

$$\begin{aligned}\hat{\theta}_{MAP} &= \operatorname*{argmax}_{\theta} p(\theta|\mathcal{D}) \\ &= \operatorname*{argmax}_{\theta} p(\mathcal{D}|\theta)p(\theta)\end{aligned}$$

If $p(\theta) \sim \mathit{Uniform}$ (a constant), MAP and MLE are the same.
Typically, when $|\mathcal{D}|$ increases, MAP $\rightarrow$ MLE that the data dominate the posteriori distribution.

MAP-examples

To be continued