Machine Learning Spring 2014
Study Guide

Paul Landes

Saturday, February 22nd, 2014

List of Equations

Mean Khan's Academy

$$\mu \triangleq \sum xP(x) = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{1}$$

Variance and Standard Deviation

Khan's Academy

$$\begin{aligned} \mathrm{Var}(X) = E[(X-\mu)^2] &\triangleq \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2 \\ \mathrm{Var}(X) = \sigma^2 &\iff \sigma = \sqrt{\mathrm{Var}(X)} \\ \iff \int (x-\mu)^2 f(x)dx &= \int x^2 f(x)dx - \mu^2 \end{aligned} \tag{2}$$

Likelihood Per Book

Page 67

$$P(\mathcal{D} \mid h) = \left[\frac{1}{|h|}\right]^N \tag{3}$$

Size principle/Occam's razor: model favors simplest hypothesis consistent with data.

Likelihood Function

Page Slides PS 2 19

$$P(\mathcal{D} \mid \theta) \tag{4}$$

Probability of data given this model hypothesis.

Maximum Likelihood Estimation (MLE)

Slides PS 2 22

$$\hat{\theta}_{\text{MLE}} = \underset{\theta}{\operatorname{argmax}}\, P(\mathcal{D} \mid \theta) \tag{5}$$

Choose parameters that make the data most probable.

Maximum a Posteriori Estimate (MAP)

Slides PS 2 37

$$\hat{\theta}_{\text{map}} = \underset{\theta}{\operatorname{argmax}}\, P(\mathcal{D} \mid \theta)P(\theta) \tag{6}$$

Somewhat Bayesian ("point + prior").

- Choose a likelihood function $P(\mathcal{D} \mid \theta)$
- Choose a prior parameter dist $P(\theta)$
- Use MAP.
- Make predictions using MAP: $P(\mathcal{D}' \mid \hat{\theta}_{\text{map}})$

Likelihood and MLE Example Office Hours

$$P(\theta) = \begin{cases} \frac{1}{3} & \theta = \theta_1 \\ \frac{1}{3} & \theta = \theta_2 \\ \frac{1}{3} & \theta = \theta_3 \end{cases}$$

$\mathcal{D} = \{H, H\}$ (two heads)
$\theta_1 = .5$
$\theta_2 = .3$
$\theta_3 = .8$
$\mathcal{L}_1 = .5 \cdot .5 = .25$

Probability of two heads (H H) using likelihood and MAP estimates:

N	$\mathcal{L}_N$	$P(\mathcal{D} \mid \theta)P(\theta)$ (MAP)
1	$.5 \cdot .5 = .25$	$.25 \cdot \frac{1}{3} = .083$
2	$.3 \cdot .3 = .09$	$.09 \cdot \frac{1}{3} = .03$
3	$.8 \cdot .8 = .64$	$.64 \cdot \frac{1}{3} = .213$

Parameter θ_3 wins with an MLE $\hat{\theta}_{\text{MLE}} = .64$.
Parameter θ_3 wins with an MAP $\hat{\theta}_{\text{map}} = .213$.

Bayes:

$$P(\theta \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \theta)P(\theta)}{P(\mathcal{D})} = \frac{P(\mathcal{D} \mid \theta)P(\theta)}{\sum_{\theta'} P(\mathcal{D} \mid \theta')P(\theta')} =$$
$$P(H \mid \mathcal{D}) = \sum_{\theta' \in \{\theta_1 \theta_2, \theta_3\}} P(\theta' \mid \mathcal{D})P(H \mid \mathcal{D}) =$$
$$\frac{.083 \cdot .5 + .03 \cdot .3 + .213 \cdot .8}{.083 + .03 + .213} = .67$$

Bernoulli/Binomial Distribution Page 34

$$\text{Bin}(k \mid n, \theta) \triangleq \binom{n}{k} \theta^k (1-\theta)^{n-k} \tag{7}$$

Models outcomes of coin tosses. Coin toss n times with $X \in \{0, \ldots, n\}$. Probability θ is the parameter of, for example, fairness of the coin.

- Expected number of "heads" outcomes in 10 flips.
- Given 10 flips what is the probability of at least 8 being "heads"?
- Given 3 heads, 2 tails, what is the estimate of θ?
- How far will this estimate be from true θ^*?

Bernoulli is the case where $n = 1$ (that is, one coin toss).

Binomial Coefficient Page 34

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{8}$$

Binomial coefficients are a family of positive integers that occur as coefficients in the binomial theorem.

Multinomial/Multinouli Distribution Page 35

$$\begin{gathered}\mathrm{Mu}(\mathbf{x} \mid n, \theta) \triangleq \binom{n}{x_1 \ldots x_K} \prod_{j=1}^{K} \theta_j^{x_j} \\ \theta_j \text{ is the probability that side } j \text{ shows up, and} \\ \binom{n}{x_1 \ldots x_K} \triangleq \frac{n!}{x_1! x_2! \ldots x_K!}\end{gathered} \tag{9}$$

Models the outcomes of tossing a K-sided die. Let $\mathbf{x} = (x_1 \ldots x_K)$ be a randome vector, where x_j is the number of times side j of the die occurs.

Poisson/Empirical Distribution Page 37

Gaussian (Normal) Distribution Page 38

$$P(x) \triangleq N(x \mid \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2 / 2\sigma^2} \tag{10}$$

Beta Distribution Page 42

$$
\begin{aligned}
\mathrm{Beta}(x \mid a, b) &= \frac{1}{B(a,b)^{x^{a-1}(1-x)^{b-1}}} \\
B &\triangleq \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}
\end{aligned} \tag{11}
$$

Where Γ is the conjugate prior.

Product Rule ES Class

$$
\begin{aligned}
P(a \wedge b) = P(a \mid b)P(b) \Leftrightarrow \\
P(X = x, Y = y) = P(X = x \mid Y = y) \cdot P(Y = y)
\end{aligned} \tag{12}
$$

Bayes Rule Page 29

$$
P(X = x \mid Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{P(X = x)P(Y = y \mid X = x)}{\sum_{x'} P(X = x')P(Y = y \mid X = x')} \tag{13}
$$