Basic ML, Bayes

Zhuoyuan Chen*
Facebook AI Research
Menlo Park, CA 94025
chenzhuoyuan07@gmail.com

Abstract

Basic machine learning

1 Probability

Quantile: cdf F, quantile F^{-1}

1.1 Discrete

Multinomial distribution:

$$Mu(x|n,\theta) = C(n; x_1, ..., x_K) \prod_{j=1}^{K} \theta_j^{x_j}$$

Poisson (to count rare events):

$$Poi(x|\lambda) = e^{-\lambda} \frac{\lambda^x}{x!}$$

Empirical distribution:

$$p_{emp}(A) = \frac{1}{N} \sum \delta_{x_i}(A)$$

1.2 Continuous

Gaussian:

$$N(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$N(x|\mu,\Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp[-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)]$$

Student-t:

$$N(x|\mu,\sigma^2,\nu) = [1 + \frac{1}{\mu}(\frac{x-\mu}{\sigma})^2]^{-(\frac{\nu+1}{2})}$$

mean: μ, model: μ, variance: $\frac{\nu\sigma^2}{\nu-2}$

Laplace:

$$Lap(x|\mu,b) = \frac{1}{2b} \exp(-\frac{|x-\mu|}{b})$$

Beta:

$$Beta(x|a,b) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}$$

*

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

1.3 Joint Distribution

covariance:

$$cov[X,Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

Dirichlet distribution (support on simplex):

$$Dir(x|\alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} x_k^{\alpha_k - 1}$$

Transformation:

$$p_y(y) = p_x(x)|detJ|$$

Central limit theorem, i.i.d. $p(x)$ with mean μ and variance σ^2.

$$Z_n = \frac{\bar{X} - \mu}{\sigma/\sqrt{N}}$$

2 Generative Modeling

Naive Bayes: assume features are independent

$$p(x|y=c,\theta) = \prod_{j=1}^{D} p(x_j|y=c,\theta_{jc})$$

3 Gaussian

Gaussian:

$$N(x|\mu,\Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)] \tag{1}$$

Theorem: N i.i.d. samples $x_i \sim N(\mu,\Sigma)$, then MLE for parameter is:

$$\mu_{mle} = \frac{1}{N}\sum_i x_i \tag{2}$$

$$\Sigma_{mle} = \frac{1}{N}(x_i - \bar{x})(x_i - \bar{x})^T \tag{3}$$

Theorem: let $q(x)$ be any density $\int q(x)x_i x_j = \Sigma_{ij}$, Then $h(q) \le h(p)$

3.1 LDA

3.2 Joint inference

$\mu = \begin{pmatrix}\mu_1 \\ \mu_2\end{pmatrix}$ $\Sigma = \begin{pmatrix}\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}$ $\Lambda = \Sigma^{-1} = \begin{pmatrix}\Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22}\end{pmatrix}$

The marginal:

$$p(x_1) = N(x_1|\mu_1,\Sigma_{11}) \tag{4}$$

$$p(x_2) = N(x_2|\mu_2,\Sigma_{22}) \tag{5}$$

Posterior:

$$
\begin{aligned}
p(x_1|x_2) &= N(x_1|\mu_{1|2},\Sigma_{1|2}) && (6)\\
\mu_{1|2} &= \mu_1 - \Lambda_{11}^{-1}\Lambda_{12}(x_2 - \mu_2) && (7)\\
&= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) && (8)\\
\Sigma_{1|2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \Lambda_{11}^{-1} && (9)
\end{aligned}
$$

2D-case: $\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$ Then

$$p(x_1|x_2) = N(\mu_1 + \frac{\rho\sigma_1}{\sigma_2}(x_2 - \mu_2), \sigma_1^2(1 - \rho^2)) \tag{10}$$