Machine Learning 2 - Cheat Sheet

Andrea Jemmett - April 16, 2016

1 Probability Theory

1.1 Independence

$p(X,Y)=p(X)p(Y) \Leftrightarrow p(X|Y)=p(X) \Leftrightarrow p(Y|X)=p(Y)$

1.2 Conditional Independence

$X \perp\!\!\!\perp Y \mid Z \iff p(X,Y|Z)=p(X|Z)p(Y|Z)$

1.3 Sum and Product Rules

$p(X,Y)=p(X)p(Y|X)$, $\quad p(X,Y,Z)=p(X)p(Y|X)p(Z|X,Y)$
$p(X)=\sum_Y p(X,Y)$

1.4 Bayes' Theorem

$p(Y|X)=\frac{p(X|Y)p(Y)}{p(X)}$, $\quad p(Y|X,Z)=\frac{p(X|Y,Z)p(Y|Z)}{p(X|Z)}$

2 Distributions

Binary	Bernoulli	Binomial	Beta
Discrete	Categorical	Multinomial	Dirichlet

2.1 Bernoulli Distribution

$\mathrm{Ber}(x|n)=\mu^x(1-\mu)^{1-x}$, $\quad \mathbb{E}[x]=\mu$, $\quad \mathrm{V}ar[x]=\mu-\mu^2$,
$P(D,\mu)=\prod_{n=1}^N \mu^{x_n}(1-\mu)^{1-x_n}$, $\quad \mu_{\mathrm{ML}}=\frac{1}{N}\sum_{n=1}^N x_n$

2.2 Binomial Distribution

$\mathrm{Bin}(m|N,\mu)=\binom{N}{m}\mu^m(1-\mu)^{N-m}$, $\quad \frac{n!}{k!(n-k)!}=\binom{n}{k}$,
$\mathbb{E}[m]=N\mu$, $\quad \mathrm{V}ar[m]=N\mu(1-\mu)$, $\quad \mu_{\mathrm{ML}}=\frac{m}{N}$

2.3 Categorical Distribution

$p(\mathbf{x}|\boldsymbol{\mu})=\prod_k \mu_k^{x_k}$, $\quad \boldsymbol{\mu}\in\{0,1\}^K$, $\quad \sum_k \mu_k=1$, $\quad \boldsymbol{\mu}_{\mathrm{ML}}=\frac{\mathbf{m}}{N}$,
$m_k=\sum_n x_{nk}$, $\quad \mathrm{Mult}(m_1\dots,m_k|N,\boldsymbol{\mu})=(\frac{N!}{m_1!,\dots,m_k}\prod_k)\mu_k^{mk}$

2.4 Beta Distribution

$\mathrm{Beta}(\mu|a,b)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$, [1] $\quad \mathbb{E}[\mu]=\frac{a}{a+b}$,
$\mathrm{V}ar[x]=\frac{ab}{(a+b)^2(a+b+1)}$, $\quad p(\mu|m,l,a,b)\propto\mu^{m+a-1}(1-\mu)^{l+b-1}$

2.5 Gamma Distribution

$\mathrm{Gamma}(\tau|a,b)=\frac{b^a}{\Gamma(a)}\tau^{a-1}e^{-b\tau}$, $\quad \mathbb{E}[\tau]=\frac{a}{b}$, $\quad \mathrm{V}ar[\tau]=\frac{a}{b^2}$,
$\mathrm{mode}[\tau]=\frac{a-1}{b}$ for $a\geq 1$, $\quad \mathbb{E}[\ln\tau]=\psi(a)-\ln b$,
$H(\tau)=ln\Gamma(a)-(a-1)\psi(a)-\ln b+a$

2.6 Multinomial Distribution

$\mathbf{x}=[0,0,0,0,1,0,0]^\mathsf{T}$, $\quad \sum_{k=1}^K x_k=1$, $\quad p(\mathbf{x}|\boldsymbol{\mu})=\prod_{k=1}^K \mu_k^{x_k}$,
$\sum_{k=1}^K \mu_k=1$, $\quad \mu_k^{\mathrm{ML}}=\frac{m_k}{N}$, $\quad m_k=\sum_{k=1}^K x_{nk}$

2.7 Dirichlet Distribution

$\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha})=\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_k)}\prod_{k=1}^K \mu_k^{\alpha_k-1}$, $\quad \alpha_0=\sum_{k=1}^K \alpha_k$

2.8 Gaussian Distribution

$\mathcal{N}(x|\mu,\sigma)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{1}{2\sigma^2}(x-\mu)^2)$,
$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})=(2\pi)^{-\frac{D}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$

2.8.1 ML for the Gaussian

$\ln p(X|\boldsymbol{\mu},\boldsymbol{\Sigma})=-\frac{ND}{2}\ln(2\pi)-\frac{N}{2}\ln|\boldsymbol{\Sigma}|-\frac{1}{2}\sum_{n=1}^N(\mathbf{x}_n-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu})$,
$\boldsymbol{\mu}_{\mathrm{ML}}=1/N\sum_{n=1}^N \mathbf{x}_n$, $\quad \boldsymbol{\Sigma}_{\mathrm{ML}}=1/N\sum_{n=1}^N(\mathbf{x}_n-\boldsymbol{\mu})^\mathsf{T}(\mathbf{x}_n-\boldsymbol{\mu})$

2.8.2 Stochastic gradient descent Gaussian

$\max\ P(x_1,\cdots,x_n|\theta)$, $\theta^N=\theta^{N-1}+\alpha_{N-1}\frac{\partial}{\partial\theta^{N-1}}\ln\ p(x_n|\theta^{N-1})$

2.8.3 Marginal and Conditional Gaussians

Given $p(\mathbf{x})=\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda}^{-1})$ and $p(\mathbf{y}|\mathbf{x})=\mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x}+\mathbf{b},\mathbf{L}^{-1})$. We get
$p(\mathbf{y})=\mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu}+\mathbf{b},\mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\mathsf{T})$ and
$p(\mathbf{x}|\mathbf{y})=\mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}[\mathbf{A}^\mathsf{T}\mathbf{L}(\mathbf{y}-\mathbf{b})+\boldsymbol{\Lambda}\boldsymbol{\mu}],\boldsymbol{\Sigma})$
where $\boldsymbol{\Sigma}=(\boldsymbol{\Lambda}+\mathbf{A}^\mathsf{T}\mathbf{L}\mathbf{A})^{-1}$.

2.9 Student's T distribution

The heavy tail of the student-t distribution makes it more robust against outliers.
$St(x|\mu,\lambda,\nu)=\frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu/2)}(\frac{\lambda^{1/2}}{(\pi\nu)^{D/2}})[1+\frac{\lambda(x-\mu)^2}{\nu}]^{-\nu/2-D/2}$,
$f_x(x)=\frac{\Gamma[(\nu+p)/2]}{\Gamma(\nu/2)\nu^{p/2}\pi^{p/2}|\Sigma|^{1/2}[1+1/\nu(x-\mu)^T\Sigma^{-1}(x-\mu)]^{(\nu+p)/2}}$
$\mathbb{E}(\mathbf{x})=\frac{\Gamma(D/2+v/2)}{\Gamma(v/2)}\frac{|\Lambda|^{1/2}}{(\pi v)^{D/2}}\int[1+\frac{(x-\mu)^T\Lambda(x-\mu)}{v}]^{-D/2-v/2}x\mathrm{d}x$

[1] $\Gamma(x)=\int_0^1 u^{x-1}e^{-u}=1$, $\quad \Gamma(x+1)=\Gamma(x)x$, $\quad \Gamma(x+1)=x!$