Prob/Stats Cheatsheet

Steve Young

ABSTRACT: Everything I know about prob/stats/maybe information theory too..

Contents

1 Conventions **1**

2 Distributions **1**

2.1 Gaussians 1

2.1.1 Basics 1

2.1.2 Differentiation moment trick 2

2.1.3 Gaussian with Linear Term 2

2.1.4 Multivariate Gaussians 3

2.2 Bernoulli 3

3 Prob and stats **3**

3.1 The Rules of Probability 3

3.2 Bayes' Rule 3

3.3 Covariance 4

4 Information Theory **4**

5 Bayesian **4**

6 Optimal Stopping Theory **4**

1 Conventions

Math Notation

2 Distributions

2.1 Gaussians

2.1.1 Basics

1. To start with, *memorize* that

$$\boxed{\int_{-\infty}^{\infty} dx\, e^{-x^2} = \pi^{1/2}} \tag{2.1}$$

2. Next, anything multiplying the x^2 in the integrand is present in inverse under the square root.

$$\int_{-\infty}^{\infty} dx\, e^{-\text{stuff}\, x^2} = \left(\frac{\pi}{\text{stuff}}\right)^{1/2} \tag{2.2}$$

so, for example:

$$\int_{-\infty}^{\infty} dx\, e^{-\frac{1}{2}ax^2} = \left(\frac{2\pi}{a}\right)^{1/2} \tag{2.3}$$

3. The traditional Gaussian pdf is thus easily seen to be

$$\mathcal{N}(x|0,\sigma^2) = \frac{1}{\left(2\pi\sigma^2\right)^{1/2}}\, e^{-\frac{1}{2\sigma^2}x^2} \tag{2.4}$$

2.1.2 Differentiation moment trick

By differentiating Eq. 2.3 wrt a, we obtain an expression for integrals of the form $\int_{-\infty}^{\infty} dx\, x^{2n} e^{-\frac{1}{2}ax^2}$, with $n \in \mathbb{Z}^+$.

e.g. for $n = 1$:

$$-2\frac{d}{da}\int_{-\infty}^{\infty} dx\, e^{-\frac{1}{2}ax^2} = \int_{-\infty}^{\infty} dx\, x^2 e^{-\frac{1}{2}ax^2} = -2\frac{d}{da}\left(\frac{2\pi}{a}\right)^{1/2} = \left(\frac{2\pi}{a}\right)^{1/2}\frac{1}{a} \tag{2.5}$$

For $n = 2$:

$$\left(-2\frac{d}{da}\right)^2\int_{-\infty}^{\infty} dx\, e^{-\frac{1}{2}ax^2} = \int_{-\infty}^{\infty} dx\, x^4 e^{-\frac{1}{2}ax^2} = \left(-2\frac{d}{da}\right)^2\left(\frac{2\pi}{a}\right)^{1/2} = \left(\frac{2\pi}{a}\right)^{1/2}\frac{1}{a}\frac{3}{a} \tag{2.6}$$

We thus obtain an expression for the expectation value of x^{2n} under the Gaussian distribution:

$$\langle x^{2n}\rangle = \frac{\int_{-\infty}^{\infty} dx\, x^{2n} e^{-\frac{1}{2}ax^2}}{\int_{-\infty}^{\infty} dx\, e^{-\frac{1}{2}ax^2}} = \frac{1}{a^n}(2n-1)(2n-3)\cdots 5\cdot 3\cdot 1 \tag{2.7}$$

2.1.3 Gaussian with Linear Term

To evaluate integrals of the form

$$\int_{-\infty}^{\infty} dx\, e^{-\frac{1}{2}ax^2+Jx}, \tag{2.8}$$

first complete the square in the exponent

$$-\frac{a}{2}x^2 + Jx = -\frac{a}{2}(x^2 - \frac{2Jx}{a}) = -\frac{a}{2}\left(x - \frac{J}{a}\right)^2 + \frac{J^2}{2a} \tag{2.9}$$

which gives

$$\int_{-\infty}^{\infty} dx\, e^{-\frac{1}{2}ax^2+Jx} = \int_{-\infty}^{\infty} dx\, e^{-\frac{1}{2}a\,(x-J/a)}e^{J^2/2a} = \left(\frac{2\pi}{a}\right)^{1/2} e^{J^2/2a} \tag{2.10}$$

where the first integral is done by shifting $x \to x + Ja$.

Eq. 2.3 is also (close to) the ***moment generating function*** of the Gaussian distribution. Given a pdf $p(x)$, the moment generating function is defined as

$$\psi_x(J) = \int_{-\infty}^{\infty} dx\, e^{Jx} p(x) \tag{2.11}$$

so that the moment generating function for the Gaussian distribution is

$$\frac{1}{(2\pi\sigma^2)^{1/2}} \int_{-\infty}^{\infty} dx\, e^{Jx} e^{-\frac{1}{2\sigma^2}x^2}, \tag{2.12}$$

TODO: Finish. Figure out clear way to include normalization factor of pdf in exposition

2.1.4 Multivariate Gaussians

Promoting a to a real $N \times N$ symmetric matrix $\mathbf{A}$, and x and J to a N-dim vectors $\vec{x}$ and $\vec{J}$ with components x_i and J_i, we have the multivariate Gaussian integral

$$\prod_{i=1}^{N} \left(\int_{-\infty}^{\infty} dx_i \right) e^{-\frac{1}{2}\vec{x}^T \mathbf{A} \vec{x} + \vec{J}^T \vec{x}} = \left(\frac{(2\pi)^N}{|\mathbf{A}|} \right)^{1/2} e^{\frac{1}{2}\vec{J}^T \mathbf{A}^{-1} \vec{J}} \tag{2.13}$$

TODO: finish —

2.2 Bernoulli

For $x \in \{0, 1\}$, Bernoulli dist parametrized by μ, with

$$p(x; \mu) = \mu^x (1 - \mu)^{1-x} \tag{2.14}$$

3 Prob and stats

3.1 The Rules of Probability

- **Product Rule**: $p(x, y) = p(x|y)p(y) = p(y|x)p(x)$
- **Sum Rule**: $p(x) = \sum_y p(x, y) = \sum_y p(x|y)p(y)$

3.2 Bayes' Rule

Using $p(y|x)p(x) = p(x, y) = p(x|y)p(y)$, we have

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} = \frac{p(x|y)p(y)}{\sum_y p(x|y)p(y)} \tag{3.1}$$

3.3 Covariance

4 Information Theory

KL divergence:

$$
\begin{aligned}
KL\big[p(x)||q(x)\big] &= \sum_{x_i} p(x_i)\log\left(\frac{p(x_i)}{q(x_i)}\right) = -\sum_{x_i} p(x_i)\log\left(\frac{q(x_i)}{p(x_i)}\right) \\
&= -\sum_{x_i} p(x_i)\log q(x_i) + \sum_{x_i} p(x_i)\log p(x_i) \\
&= H(p,q) - H(p) \qquad (4.1)
\end{aligned}
$$

where $H(p,q)$ is the cross entropy, and $H(p)$ is the entropy.

5 Bayesian

6 Optimal Stopping Theory