February 18, 2010

1 Probability

1.1 Covariance

$$\begin{aligned} Var(X) &= \mathbb{E}[(X - \mathbb{E}(X))^2] \\ Cov(X,Y) &= \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \end{aligned}$$

If X and Y are uncorrelated ($Cov(X,Y) = 0$), they are not necessarily independent.

1.2 Markov Inequality

Consider a nonnegative random variable X,

$$\mathbb{P}(X > a) \leq \frac{\mathbb{E}(X)}{a}.$$

The proof follows from observing $1(X > a) \leq \frac{X}{a}$, and taking expectations on each side.

1.3 Chebyshev Inequality

$$\mathbb{P}(|X - \mathbb{E}(X)| > a) \leq \frac{Var(X)}{a^2}$$

This is proved by using Markov's Inequality with the nonnegative random variable $(X - \mathbb{E}(X))^2$.

Equivalently, replacing a with $\alpha\,\mathrm{sdev}(X)$ we can write,

$$\mathbb{P}(|X - \mathbb{E}(X)| > \alpha\,\mathrm{sdev}(X)) \leq \frac{1}{\alpha^2}.$$

1.4 Chernoff Bounding

Apply Markov property to $e^{t[X - \mathbb{E}(\mathbb{X})]}$.

1.5 Variance and Covariance

Let $X \in \mathcal{R}^d$, define the covariance matrix,

$$\Sigma = \mathbb{E}([X - \mathbb{E}(X)][X - \mathbb{E}(X)]^T). \tag{1}$$

Next apply the Markov Inequality to $Z = [X - \mathbb{E}(X)]^T \Sigma^{-1} [X - \mathbb{E}(X)]$,

$$\begin{aligned}
\mathbb{P}(Z > a) &\leq \frac{\mathbb{E}(Z)}{a} \\
&= \frac{\mathbb{E}([X - \mathbb{E}(X)]^T \Sigma^{-1} [X - \mathbb{E}(X)])}{a} \\
&= \frac{\mathbb{E}(trace([X - \mathbb{E}(X)]^T \Sigma^{-1} [X - \mathbb{E}(X)]))}{a} \\
&= \frac{\mathbb{E}(trace([X - \mathbb{E}(X)][X - \mathbb{E}(X)]^T \Sigma^{-1}))}{a} \\
&= \frac{trace(\Sigma\Sigma^{-1})}{a} \\
&= \frac{trace(I_d)}{a} \\
&= \frac{d}{a}
\end{aligned}$$

Variance and Covariance are indicators of linear dependence.

1.6 Law of Large Numbers

Let $X_1, ..., X_n$ be independent and $\mathbb{E}(X_i) = \mu$ and $Var(X_i) = \sigma^2$. Define $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} (X_i)$. Then $\mathbb{E}(\bar{X}) = \mu$ and

$$\begin{aligned}
Var(\bar{X}) &= \frac{1}{n^2} \sum_{i=1}^{n} Var(X_i) \\
&= \frac{1}{n^2} \sum_{i=1}^{n} \sigma^2 \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

Now applying the Chebyshev Inequality,

$$\mathbb{P}(|\bar{X} - \mu| > a) \leq \frac{\sigma^2}{na}.$$

1.7 Probability Definitions

1.7.1 Probability Measures

The paradox of the great circle motivates careful definitions of probabilities. Ω is the set of outcomes, and $\mathbb{P}(\Omega) = 1$. If $A \cap B = \emptyset$, then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$. Another property of a probability measure is countable additivity: if $A_1, A_2, ...$ are disjointed, $\mathbb{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$. Futhermore, we only take probabilities of events that are Borel Sets.

1.7.2 Cumulative Distribution Function (CDF)

The cdf is defined, $F(x) = \mathbb{P}(X \leq x)$. Note $\mathbb{P}(X \in (a, b]) = F(b) - F(a)$. If $X \in \mathcal{R}^d$, $F(x) = \mathbb{P}(X_1 \leq x_1, ..., X_d \leq x_d)$.

1.7.3 Density Function (PDF)

If F is differentiable, the density is defined, $p(x) = F'(x)$. We can write $\mathbb{P}(X \in (a, b]) = \int_{x \in (a,b]} p(x)dx$. The expected value of X can be calculated with $\mathbb{E}(X) = \int_{-\infty}^{\infty} xp(x)dx$. The total area under the density function is 1, $\int_{-\infty}^{\infty} p(x)dx = 1$. Note that it is usually much harder to estimate the density function than the cumulative distribution function.

1.7.4 The Normal Distribution

Let Z be normally distributed with mean 0 and variance 1, $Z \sim \mathcal{N}(0, 1)$. The pdf can be written,

$$p(z) = \phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}.$$

The cdf can be written,

$$F(z) = \Phi(z) = \frac{1}{2}[1 + erf(\frac{z}{\sqrt{2}})]. \tag{2}$$

Now let X be normally distributed with mean μ and variance σ^2, $X \sim \mathcal{N}(\mu, \sigma^2)$. The pdf can be written,

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

The cdf can be written,

$$F(x) = \Phi(\frac{x - \mu}{\sigma}).$$

1.8 Central Limit Theorem

Let $X_1, ..., X_n$ be independent with mean μ and variance σ^2. Then approximately

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma} \sim \mathcal{N}(0,1).$$

The central limit theorem states as n goes to ∞,

$$\mathbb{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma} > a \right) \to 1 - \Phi(a). \tag{3}$$

1.9 Law of Large Numbers

$$\mathbb{P}(|\bar{X} - \mu| > a) \leq \frac{1}{na^2} \to 0,$$

as n goes to ∞.

2 Comparing Classifiers

The goal is to statistically compare the performance of classifiers C_1 and C_2. Define

$$R_i = \begin{cases} +1 & \text{if } C_2 \text{ correct and } C_1 \text{ incorrect,} \\ 0 & \text{if they agree,} \\ -1 & \text{If } C_1 \text{ correct and } C_2 \text{ incorrect.} \end{cases}$$

Assume the R_i's are independent and $\mathbb{E}(R_i) = \mu$ and $Var(R_i) = \sigma^2$. Also define $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} r_i$ and $\bar{R}_n = \frac{1}{n}\sum_{i=1}^{n} R_i$. If R_i is significantly greater than 0, C_2 is better. If R_i is significantly smaller than 0, C_1 is better.

2.1 Central Limit Theorem

We see $\mathbb{E}(\bar{R}_n) = \mu$ and $Std(\bar{R}_n) = \frac{\sigma}{\sqrt{n}}$. By the CLT,

$$\frac{\bar{R}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1).$$

If C_2 is worse than C_1, $\mathbb{P}(\bar{R}_n > \hat{\mu}) \approx 1 - \Phi(\frac{\hat{\mu}-\mu}{\sigma}\sqrt{n}) \leq 1 - \Phi\left(\frac{\hat{\mu}}{\sigma}\sqrt{n}\right) = \Phi\left(-\frac{\hat{\mu}}{\sigma}\sqrt{n}\right)$

2.2 Student's t-distribution

Define $\hat{\sigma} = \frac{1}{n-1}\sum_{i=1}^{n}(r_i - \hat{\mu})^2$ and $\bar{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(R_i - \bar{R})^2$. Then

$$\frac{\bar{R}_n - \nu}{S_n/\sqrt{n}} \sim \text{student's t-distribution with n-1 degrees of freedom}$$

2.3 Chernoff Bounding

$\mathbb{P}(\bar{R}_n > \hat{\mu}) \leq e^{-\frac{n^2\hat{\mu}^2}{r}}$