Scribe to lecture Tuesday March 16 2004

Scribe outlines:
Message
Confidence intervals
Central limit theorem
Em-algorithm
Bayesian versus classical statistic

Note: There is no scribe for the beginning of the lecture Thursday March 19. The missing part of the lecture focuses on the EM algorithm and clustering.

Message

Left to do is Support Vector Machines, reinforcement learning and Inductive logic programming. We don't have time to do both SVM and ILP so we have to choose between:

1. Skip support vector machines. 12 votes
2. SVM-light. Spend one class on it. 1 vote
3. Spend two classes on SVM. 4 votes

Confidence intervals

Our goal is to estimate a parameter value e.g. the mean μ given the variance σ. By identifying a constant c you can find an interval such as the probability of getting a data point d within the interval $(\mu - c, \mu + c)$ is, for example $\geq 95\%$. $P(\mu - c < d < \mu + c) \geq 95\%$. Used the other way around you can find an interval to which μ belongs with 95% confidence.

Procedure
1. Observe a data point d.
2. Output: $\mu \in (d - c , d + c)$ with confidence 95%.
   $\mu - c < d \Leftrightarrow \mu < d + c$
   $\mu + c > d \Leftrightarrow \mu > d-c$
   $d-c < \mu < d+c$

So what does this mean? No matter what the true value of μ is the probability that $\mu \in$ the output interval is $\geq 95\%$. For example $P(-0.5 < \mu < 0.5) \geq 95\%$. The data point d is a random variable. Note that μ isn't a random variable, it's fix but unknown. This is why you call it confidence interval and not probability interval.

To see the difference in the unknown parameter μ and the random variable d think of the following. Assign all people in the class to a drug experiment. You would have a high probability of getting as many women as men in a group (not bad for a computing class!). This corresponds to the mean μ. But this is different from looking at an actual group and see if there are as many men as women. This would correspond to a sample d.

Confidence interval for a uniform distribution

Take the parameter θ to be fixed but unknown. Let D be a uniform distribution over the interval $(\theta-0.5,\theta+0.5)$. (In a uniform distribution every point have the same probability of being selected. The graph is box shaped compared to the bell shaped normal distribution where the probability is higher near the mean.) Let d be a sample from the interval.
The probability of getting a value for d within an interval of length 0.8 centered around θ is 80% since $(\theta-0.4,\theta+0.4)$ is 80% of the total interval $(\theta-0.5,\theta+0.5)$.
$P(d \in (\theta-0.4,\theta+0.4))=0.8$. For this distribution use $c=c_{0.8}=0.4$ for 80% confidence interval.

You can also use this knowledge to go the other way around and draw conclusions of θ based on the value of the sample d you get. If d = 0.7 this lets you conclude that $\theta \in (0.7-0.4,0.7+0.4)= (0.3,1.1)$ with 80% confidence or $\theta \in (0.7-0.25,0.7+0.25)= (0.45,0.95)$ with 50% confidence ($c_{0.5}=0.25$).

Now we take two random samples d1=0.3 and d2=0.9.

- Looking at the data points we can reach some conclusions about the value of θ. The first sample gives us $\theta \in (0.3-0.5,0.3+0.5) =(-0.2,0.8)$ with 100% confidence and the second gives us $\theta \in (0.9-0.5,0.9+0.5) = (0.4,1.4)$ with 100% confidence. Together we have that $\theta \in (0.4, 0.8)$ with 100% confidence.
- Another way of doing this without having to look at the data points is to use the procedure: $P(\min(d1,d2)-c_a< \theta <\max(d1,d2)+c_a) \geq a$.
  $P(\min(d1d2)-c_{80}<\theta<\max(d1,d2)+ c_{80}) = P(0.9-0.4< \theta<0.3+0.4) = P(0.5< \theta<0.7) \geq 80\%$
  $P(\min(d1,d2)-c_{100}<\theta<\max(d1,d2)+c_{100}) = P(0.9-0.5<\theta<0.3+0.5) = P(0.4<\theta<0.8)=100\%$

Confidence intervals versus posterior

- A confidence interval is not the same as posterior probability.
- From a Bayesian point of view, given the observation Y=y, we may be able to give a much higher probability that the desired mean μ lies in the confidence interval.

For all sorts of priors $P(\text{Theta}=\theta)$ you can get $P(d1,d2|\text{Theta}=\theta)$, where d1 and d2 are the values of two random samples. Using Bayes' formula you can calculate the posterior $P(\text{Theta}=\theta|d1,d2)$. And get an interval for $P(\text{Theta}=\theta|d1,d2)$. In this way Bayes probability is close to classical confidence intervals.

Confidence interval for other distributions

Confidence intervals can be used to evaluate how good a learned hypothesis is. You can do this by letting the hypothesis classify n random instances. The number of errors committed r, will be distributed with the Binomial distribution. If p is be the probability for the hypothesis to misclassify an instance than E[x]=np and $\sigma_D= \sqrt{(p(1-p)n)}$, where $x=\{x1+\ldots+xn\}$.

The actual number of misclassifications of a test case with n instances is r. Use r/n as an estimator of p and r as an estimator of the sample mean. To get the sample variance you divide by the number of samples, giving us the following estimate of the variance of sample variance $\sigma_s= \sigma_D/n = \sqrt{(p(1-p)/n)}$.

To get a confidence interval you approximate the binomial distribution by a normal distribution (using the central limit theorem) with μ = sample mean and σ = sample derivation. Now it's possible to calculate the confidence interval using $N(\mu,\sigma)$.

Central Limit Theorem

If

- X1, X2, …, Xn are n random samples from a distribution
- **X** is the sample mean
- Set E[x]=E[**X**]
- Set Var(x) = n*Var(**X**)
- Let the number of samples $n \to \infty$.

Then the Central Limit theorem states that no matter which distribution you start with the normal distribution N(E[X],Var(X)) will bee an arbitrary good approximation.

Depending on the original distribution the number of samples needed before the normal distribution is a good approximation varies. For a binomial distribution it's usually enough with n=30.

This can cause problems when you have large samples. You're expected to get a very precise result, which isn't always the case.

The EM-algorithm

The EM-algorithm (section 6.12 in Mitchell's Machine learning) is used to find a maximum likelihood hypothesis. You want to find the setting that makes the observed variable as likely as possible. It can be used in many settings where we wish to estimate some set of parameters that describe an underlying probability distribution. This procedure can deal with unobserved variables. Another advantage is that it's guaranteed to find a minimum if there is any.

You could say that the general idea behind EM is that you pretend to know the parameters of the model and then interfere the probability that each data point belongs to each component. After that we refit the components to the data, where each component is fitted to the entire data set with each point weighted by the probability that it belongs to that component.

General statement

For m independent instances let $X=\{x_1, \dots , x_m\}$ denote the observed data and $Z=\{z_1, \dots , z_m\}$ the unobserved data. The full data X U Z is denoted by Y. Our current estimate of the parameters we want to estimate is denoted θ^n.

Two main steps make up the EM-algorithm: estimation and maximisation. In the estimation part the current hypothesis θ^n together with the data point x is used to estimate the unobserved variables. Next this estimation together with the data point x is used to perform a maximisation over new hypotheses θ. The best one θ^{n+1} is chosen to replace the old one.

The EM formula: $\theta^{(n+1)} = \text{argmax}_{\theta} \Sigma_z [P(Z=z|x,\theta^n) \ln(P(x,Z=z|\theta)]$.

Where $P(Z=zij|xi,\theta^n)$ = constant * $P(xi|Z=zij,\theta^n)$ * $P(Z=zij|\theta^n)$ (Bayes theorem)

When Q is a continuous function, the EM algorithm will converge to a stationary point of the likelihood function $P(y|\theta')$. A disadvantage with the EM algorithm is that you can get stuck in a local minimum. The more dimensions you have the more likely you are to find a local minimum and have a dimensional collapse.

In the example below fixed and equal variances are assumed. However the algorithm can deal with different or unknown variances. You theoretically you could run it for an unknown number of distributions, but it does helps a lot to specify the number of unknown variables. If you don't there's a big risk of overfitting.

Example of the use of the EM-algorithm

To illustrate the use of the EM–algorithm we take a look at an example where a sample X is drawn for one of two Gaussian distributions with unknown means and the same variance $N1(\mu1, \sigma)$ and $N2(\mu2, \sigma)$. Our goal is to find estimations for $\mu1$ and $\mu2$. (In Mitchell's Machine learning they assume that the probability of a getting a sample from distribution one is ½ or that $P(zi1|\theta^n) = P(zi2|\theta^n)$, this assumption isn't necessary. If you assume that the samples are drawn at random you can skip the part $P(Z=z|x,\theta^n)$.)

Using the convention above X is the observed variable and Z denotes the hidden variables, in this case Z1=1 and Z2=0 if x is from distribution number one. θ is the set of parameters to be estimated. Our current hypothesis is $\theta = <\mu1,\mu2>$.

Estimation: $P(Z=zij|xi,\theta^n) = P(xi|zij,\theta^n) * P(zij|\theta^n)$ alpha?? alpha * $P(xi,|zij|\theta^n) * P(zij|\theta^n)$
$P(xi,Z=zij|\theta)=1/\sqrt{(2*\pi*\sigma^2)} * e^{-1/(2\sigma^2)\ \Sigma j\ zij(xi-\mu j)^2}$

Maximisation: $\theta^{(n+1)} = \text{argmax}_{\theta} (\Sigma [P(Z=z|x,\theta^n) \ln(P(x,Z=z|\theta)])$

If we now assume that we know that the sample is drawn for distribution one this equation will reduce to the probability of getting a specific sample from an ordinary normal distribution:
$P(x1|Z=z11,\theta) = P(X=x1)_{N1(\mu1,\ \sigma)} = 1/\sqrt{(2*\pi*\sigma^2)} * e^{-(x1-\mu1)^2/(2\sigma^2)}$

Bayesian versus classical statistic

Pro Bayesian statistic

- Unified
- Intuitively correct
- Incremental, online
- No maximum likelihood maximisation
- Distribution free

Pro Classical statistic

- Avoid subjective priors (ok with base-ratios but no wild guesses)

- Less computational expensive than Bayesian statistic
- Much knowledge about the distribution

Lina Björnheden