CS532c Fall 2004: Homework 5
Out Fri Oct 22, Due Mon Nov 1

1 Viterbi algorithm

[40 points]

Implement the viterbi algorithm for HMMs for finding the most probable path. (If you do not understand this sentence, you should read the tutorial on HMMs by Rabiner on the course web page.) The algorithm should have the following interface:

```
function path = viterbi(prior, transmat, obsmat, data)
```

where prior(i) = $P(X_1 = i)$ is the initial state probability distribution, transmat(i,j) = $P(X_t = j|X_{t-1} = i)$ is the state transition matrix, obsmat(i,o) = $P(Y_t = o|X_t = i)$ is the observation matrix, and data (a vector) is the observation sequence. (Assume that data(t) is a symbol numbered $1, 2, \ldots, O$, where O is the size of the alphabet). The output should be a vector where path(t) is the most probable state at time t.

We will now test your algorithm by applying it to the problem of segmenting a sentence which is a mixture of German and Spanish. Load the file 'segment.mat' as follows:

```
data = load('segment.mat');
```

This has two fields: `data.gerspa` is a vector of integers representing letters. To view this, type

```
stream2text(data.gerspa)
```

The second field is `data.gerspa-lang` , which is a vector of 1s and 2s, representing the true segmentation (2 = german, 1 = spanish). Load a pre-trained HMM:

```
load('hmm.mat');
```

This has fields `hmm.prior`, `hmm.transmat` and `hmm.obsmat`. Apply this HMM and your Viterbi algorithm to `data.gerspa` and plot the estimated segmentation versus the true segmentation. How many classification errors do you make?

2 Max likelihood estimation for 1D Gaussians

[4 points per question except Q6 which gets 9 points, so total is 45.]

Recall that a univariate Gaussian (or normal) random variable, with mean μ and variance σ^2, is given by the following probability density function:

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

1. Write down the likelihood $L(x_1, \ldots, x_n; \mu, \sigma)$ of a sample drawn independently from a normal distribution with (unknown) mean and variance.

The maximum likelihood estimator $\hat{\mu}(x_1, \ldots, x_n)$, is the value of μ that maximizes the likelihood L:

$$\hat{\mu}(x_1, \ldots, x_n) = \arg\max_{\mu} \max_{\sigma} \mathrm{L}(\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{n}; \mu, \sigma)$$

Instead of searching for the maximum of *L*, we will search for the maximum of log *L*. This is fine since the logarithm is a monotonically increasing function. To find the maximum, we would like to solve the equation:

$$\frac{\partial \mathrm{logL}(\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{n}; \mu, \sigma)}{\partial \mu} = 0$$

2. Calculate $\frac{\partial \mathrm{logL}(\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{n}; \mu, \sigma)}{\partial \mu}$ and solve the above equation, in order to find the maximum likelihood estimator $\hat{\mu}$. Show that the solution does not depend on σ.

In general, we might have needed to find the values of σ which maximize L together with μ. This is luckily unnecessary, since as you showed, $\arg\max_{\mu} \mathrm{L}(\mu, \sigma)$ is independent of σ.

Note that $\hat{\mu}$ is a function of the sampled values, and thus $\hat{\mu}$ can itself be viewed as a random variable. An estimator such as $\hat{\mu}$ is said to be unbiased if the expected value of this random variable is equal to the "true" value being estimated, that is if $\mathbf{E}_{X1,\ldots,X_n\ N(\mu,\sigma)}[\hat{\mu}(X_1, \ldots, X_n)] = \mu$ for all μ, σ. The expectation here is over the possible choices of the random samples assuming they came from a Gaussian with mean μ and variance σ^2.

3. Calculate $\mathbf{E}_{X1,\ldots,X_n \sim N(\mu,\sigma)}[\hat{\mu}(X_1, \ldots, X_n)]$. Is $\hat{\mu}$ unbiased? Hint: the expectation of a sum is equal to the sum of the expectations.

We now proceed to calculate the maximum likelihood estimator for σ:

$$\hat{\sigma}(x_1, \ldots, x_n) = \arg\max_{\sigma} \max_{\mu} \mathrm{L}(\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{n}; \mu, \sigma)$$

We do so in a similar way, by taking the derivative of $\max_{\mu} \log L(x_1, \ldots, x_n; \mu, \sigma)$, with respect to σ . Note that in taking this derivative, we assume that μ is set to its maximum likelihood value. However, we already know the value of μ that maximizes $L(\mu, \sigma)$ and so can just plug it in.

4. Does it matter if we take the derivative with respect to the variance σ^2, or its square root σ?

5. Calculate $\hat{\sigma}(x_1, \ldots, x_n)$ by maximizing the log-likelihood.

6. We would now like to show that $\hat{\sigma}^2$ is not an unbiased estimator of σ^2. Calculate $\mathbf{E}_{\mathrm{X}_1,\ldots,\mathrm{X}_\mathrm{n} \sim \mathcal{N}(\mu,\sigma)}[\hat{\sigma}^2(\mathrm{X}_1, \ldots, \mathrm{X}_\mathrm{n})$ to do so. Hint: note that $X_1, \ldots, X_n$ are independent, and use the fact that the expectation of a product of independent random variables is the product of the expectations. Also, recall that for any random variable R we have $var[R] = E[E^2] - (E[R])^2$.

7. Suggest an unbiased estimator $\tilde{\sigma}^2(x_1, \ldots, x_n)$ for σ^2, based on the the maximum likelihood estimator above, and show that $\tilde{\sigma}^2$ is in fact unbiased. Hint: scale the maximum likelihood estimator so that it will be unbiased.

8. Consider a sample $x_1, \ldots, x_n$ drawn from a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$, where the true mean μ is known, but the variance is not. What is the maximum likelihood estimator for the variance in this case ? Is it unbiased ?

We now return to the case in which neither the mean nor the variance are known.

9. An estimator being unbiased does not necessarily make it good. For example, consider the following estimator for the mean of a Gaussian random variable: $\breve{\mu}(x_1, \ldots, x_n) = x_1$. Show that this is an unbiased estimator of μ.

   One reason that $\breve{\mu}$ is not a very good estimator, is that no matter how many samples we have, it will not improve. It will never converge to the true value of μ. An estimator $\hat{\theta}$ is (mean squared) *consistent* if it converges to θ in the following sense: $E_{X_1,\ldots,X_n \sim \mathcal{N}(\mu,\sigma)}[(\hat{\theta}(X_1, \ldots, X_n) - \theta)^2] \to 0$ as $n \to \infty$. In other words, the more data points we get, the less likely it is that the estimate $\hat{\theta}(X_1, \ldots, X_n)$ deviates much from θ.

10. Show that $\hat{\mu}$ (the maximum likelihood estimate of the mean) is a consistent estimator of μ.

11. (Optional - no points). Do you think $\hat{\sigma}^2$ is a consistent estimator of σ^2? What about $\tilde{\sigma}^2$?

3 MAP estimation for 1D Gaussians

[5 points per question]

The maximum a-posteriori (MAP) estimator is defined as the value of the parameters θ that maximize

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(\theta|data) = \arg\max_{\theta} p(data|\theta)p(\theta)$$

Consider samples $x1, \ldots, x_n$ from a Gaussian random variable with known variance σ^2 and unknown mean μ. We further assume a prior distribution (also Gaussian) over the mean, $\mu \sim \mathcal{N}(m, s^2)$, with fixed mean m and variance s^2.

1. Calculate the MAP estimate $\hat{\mu}_{MAP}$. Hint: as we did before, set the derivative of the logarithm to zero.

2. Show that as the number of samples increase, the prior knowledge becomes insignificant. That is, all MAP estimates assuming as a prior on μ any Gaussian distribution with non-zero variance, will converge to each other. What is the common estimator that all such MAP estimators converge to ? (Further note: This actually holds with rather mild assumptions about the prior— it need not be Gaussian).

3. What does the MAP estimator converge to if we increase the prior variance s^2?