1. 次元削減は、訓練インスタンスが莫大な特徴量を持つ場合に、訓練の高速化やデータの可視化、さらに過学習のリスクを減らすために行われる。また、これの欠点としては、この後の訓練アルゴリズムの性能が下がることやCPUに負荷がかかることがあること、パイプラインの複雑化、変換後の特徴量が解釈しにくいことがあげられる。
2. 次元の呪いとは、特徴量を高次元にしていくとデータセットが疎になる可能性が増加し、過学習しやすくなってかえって予測の精度が落ちるというもの。
3. 次元削減すると、必ずいくらかの情報が失われることになるため、元に戻すことはできない。
4. カーネルPCAという手法を用いることで、高次非線形データセットの次元削減ができる。
5. およそ200次元(元データの20%)になると予想される。
6. 逐次学習型PCAは訓練セット全体がメモリにおさまらない程大きい場合、ランダム化PCAは削減後の次元が元の次元よりもかなり小さい場合、カーネルPCAはデータセットに複雑な非線形射影を行って次元削減したい場合に有効である。これ以外では普通のPCAを用いればよい。
7. 再構築誤差の小ささで評価する。ただし、kPCAでは再構築プレイメージによって元のインスタンスとの二乗距離を計算し、その小ささで評価するといった工夫が必要である。
8. 分からない。しっかりとした意味がある。例えば、PCAで不要な次元を手っ取り早く大量に取り除いてから、LLEなどの時間のかかる次元削減アルゴリズムを使うことはよくある。この2ステップアプローチは、LLE単独の場合にほぼ匹敵する性能を引き出せるが、処理時間は数分の1である。
9. ファイル「code_8_9.ipynb」参照のこと。まず、n_estimators=500、max_leaf_nodes=16、n_jobs=-1という条件でランダムフォレストを訓練した。訓練にかかった時間は49秒で、テストセットでの正解率は83.02%であった。次に、次元削減後は、訓練時間が3分で、テストセットの正解率が81.22%であった。訓練時間がなぜか3倍になり、正解率は若干減少した。
10. ファイル「code_8_10.ipynb」参照のこと。プロットは以下のようになった。t-SNEは2次元射影がうまくいっている一方で、PCAでは2次元に特徴量を減らすと因子寄与率がかなり小さくなってしまうことが分かる。

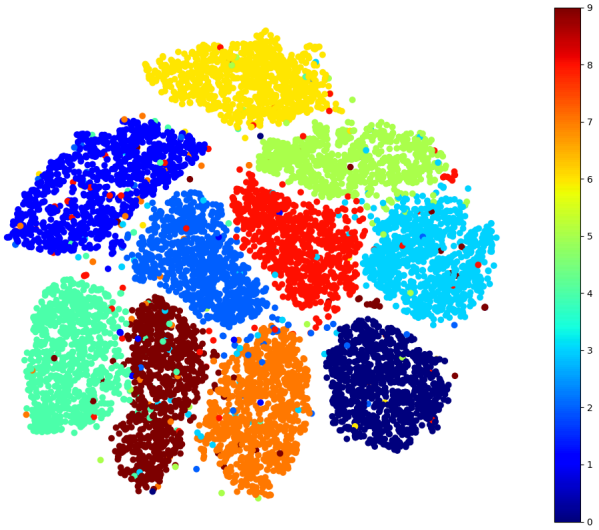

図1. t-SNEにより特徴量を2次元化したマップ

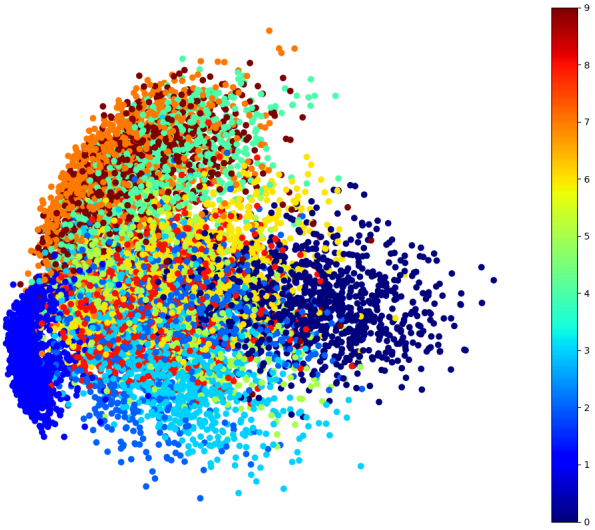

図2. PCAにより特徴量を2次元化したマップ