10-701 Machine Learning, Fall 2012: Homework 4 Solutions

1 Graphical models 2 (Zeyu, 25 points)

1.1

1. Proof: $a \perp (b,c)|d \Rightarrow p(a,b,c|d) = p(a|d)p(b,c|d)$

   $\Rightarrow \sum_c p(a,b,c|d) = \sum_c p(a|d)p(b,c|d) = p(a|d)\sum_c p(b,c|d)$

   $\Rightarrow p(a,b|d) = p(a|d)p(b|d)$

2. Assume x is a node in the graph. Consider the following three cases:

   - Paths going through the parent of x:

     $\boxed{A} \rightarrow \boxed{c} \rightarrow \boxed{\text{x}}$: head-to-tail, A and x are d-separated by c

     $\boxed{A} \leftarrow \boxed{c} \rightarrow \boxed{\text{x}}$: tail-to-tail, A and x are d-separated by c

   - Paths pointing to the parent of the children of X:

     $\boxed{A} \leftarrow \boxed{c} \leftarrow \boxed{\text{x}}$: head-to-tail, A and x are d-separated by c

   - Paths going through the co-parent (node e) of the children of X:

     $\boxed{A} \leftarrow \boxed{e} \rightarrow \boxed{c} \leftarrow \boxed{\text{x}}$: tail-to-tail at e, A and x are d-separated by e

     $\boxed{A} \rightarrow \boxed{e} \rightarrow \boxed{c} \leftarrow \boxed{\text{x}}$: head-to-tail at e, A and x are d-separated by e

3. (a) Yes.

   (b) No: $X \perp Y$ vs. $X \perp Y|A$

   (c) No: $X \perp Y|(A,B)$ vs. $X \perp Y|A$, $A \perp B|X$ vs. $A \perp B|(X,Y)$

   (d) No: $X \perp Y|B$ vs. $X \perp Y|A$

1.2 Exact Inference

1. (a) $m_{hc}(v_i) = p(h = v_i)$

   (b) $m_{ca}(v_i) = p(c = v_i) = \sum_{g,h} p(c = v_i|g,h)p(g)p(h) = \sum_{g,h} p(c = v_i|g,h)m_{gc}(g)m_{hc}(h)$

   (c) $m_{ar}(v_i) = p(a = v_i) = \sum_{b,c,d} p(a = v_i|b,c,d)m_{ba}(b)m_{ca}(c)m_{da}(d)$

2. (a) $p(r = v_i) = \sum_a p(r = v_i|a)m_{ar}(a)$

(b)

$$p(g = v_i | r = v_j) = \frac{p(g = v_i, r = v_j)}{\sum_{v_i} p(g = v_i, r = v_j)}$$

$$p(g = v_i, r = v_j) = \sum_{a,b,c,d,h} p(r|a)p(a|b,c,d)m_{ba}(b)p(d)p(c|g,h)p(g)p(h)$$

1.3 Stochastic Inference

1. Markov Blanket: A,B,C,D,E,F
2. Calculating $p(X|MB(X))$ can be efficient using variable elimination

$$p(X|MB(X)) = \frac{p(X, MB(X))}{\sum_x p(X = x, MB(X))}$$

$p(X = x, MB(X)) = p(A|B, x)p(C|x, D)p(x|E, F) \times \sum_g p(g)p(D|g) \times \sum_j p(j)p(B|j)$

Finally,

$$p(X|MB(X)) = \frac{p(A|B, x)p(C|x, D)p(x|E, F)}{\sum_x p(A|B, x)p(C|x, D)p(x|E, F)}$$

Since calculating $X|MB(X)$ is very efficient (only one-level summation), it reduces the complexity of Gibbs sampling to $O(D^2 + n)$ where D is the number of nodes; and n is the number of samples.

3. Open question
   (a) Just fix $G = 1$ when calculating $p(X = x, MB(X))$
   (b) Use weighted sampling

2 Learning Theory [20 points, Martin]

2.1 VC dimension [10 points]

Recall that, given a hypothesis space H defined over an instance space X, the Vapnik–Chervonenkis dimension VC(H) is defined to be the largest integer such that there exists a subset $x_1, ..., x_{\mathrm{VC}(H)} \in X$ that is shattered by H (i.e. for any binary labeling of $x_1, ..., x_{\mathrm{VC}(H)}$ there exists a hypothesis $h \in H$ that is consistent with that labeling). Thus, in order to prove that the VC dimension of a hypothesis space H is some integer d, we must prove both that there exists a subset of X of size d that is shattered by H, *and* that for any $D > d$ there exist **no** subsets of X of size D that are shattered by H. The latter task can seem daunting – not only do we have to prove that **all** subsets of X of size D are **not** shattered by H, but we must also prove this for all $D > d$. The following result can make this much easier:

(a) **[5 points]** Prove that if there exists a subset of X of size d (for some integer d) that is shattered by H, then for any $1 \leq k < d$ there also exists a subset of size k of X that is shattered by H.

Answer: Suppose $x_1, ..., x_d \in X$ are shattered by H. Consider the set of points $x_1, ..., x_k$. Let $l_1, ..., l_{d-1} \in \{0, 1\}$ be any labeling of those points. Let $l_{k+1}, ..., l_d = 0$. Since the d original points are shattered by H, there exists $h \in H$ such that $h(x_i) = l_i$ for $i = 1, ..., d$. In particular, the same holds true if we only consider $i = 1, ..., k$. Hence $x_1, ..., x_k$ are shattered by H.

The above claim implies that if no subset of size d is shattered, then no subset of size $D > d$ is shattered either. Hence, to prove that the VC dimension of H is d, it is sufficient to find *some* subset of X of size d that is shattered, and to prove that *no* subset of size $d+1$ is shattered.

(b) **[5 points]** Let $X = \mathbb{R}$ (one dimension). Let H be the set of all classifiers h that, for some set of non-intersecting intervals $R_1, ..., R_p$, classify a point x as $h(x) = 1$ if $x \in \bigcup_{i=1}^{p} R_i$, and $h(x) = 0$ otherwise (p is fixed and given). Find $\text{VC}(H)$ (prove your answer is correct).

Answer: Consider any set of points $x_1 < ... < x_{2p}$. Each pair of points x_{2i-1}, x_{2i} can be shattered with the i'th interval for $i = 1, ..., p$, without affecting the rest. One possible way of doing this is to use $R_i = (x_{2i-1}, x_{2i})$, $R_i = [x_{2i-1}, x_{2i})$, $R_i = (x_{2i-1}, x_{2i}]$, and $R_i = [x_{2i-1}, x_{2i}]$. So $x_1, ..., x_{2p}$ can be shattered.

Now consider $x_1 < ... < x_{2p+1}$ (note that any set of $2p+1$ points can be reordered in this form, so we have not lost generality). Label the odd-indexed points 1, and the rest 0. It is quite easy to verify that no classifier in H can be consistent with this labeling.

So, $\text{VC}(H) = 2p$.

2.2 Structural risk minimization [10 points]

Recall the PAC bound using VC dimension: given a hypothesis class H and m training samples, with probability $\geq 1 - \delta$, for all $h \in H$

$$|\text{error}_{\text{true}}(h) - \text{error}_{\text{train}}(h)| \leq \varepsilon(H, m, \delta)$$

where

$$\varepsilon(H, m, \delta) = 8\sqrt{\frac{\text{VC}(H)\left(\ln \frac{m}{\text{VC}(H)} + 1\right) + \ln \frac{8}{\delta}}{2m}}.$$

In class you saw how we can use this to bound the true error of the empirical risk minimizer $\widehat{h} = \text{argmin}_{h \in H} \text{error}_{\text{train}}(h)$; with probability $\geq 1 - \delta$,

$$\begin{aligned}
\text{error}_{\text{true}}(\widehat{h}) &\leq \text{error}_{\text{train}}(\widehat{h}) + \varepsilon(H, m, \delta) \\
&\leq \text{error}_{\text{train}}(h^*) + \varepsilon(H, m, \delta) \\
&\leq \text{error}_{\text{true}}(h^*) + 2\varepsilon(H, m, \delta)
\end{aligned}$$

where $h^* = \text{argmin}_{h \in H} \text{error}_{\text{true}}(h)$ is the true risk minimizer.

Given a set of hypothesis classes $H_1, H_2, ..., H_K$ (K possibly infinite) with $\text{VC}(H_1) \leq \text{VC}(H_2) \leq ... \leq \text{VC}(H_K)$, *structural risk minimization* is the following procedure. First for each $k = 1, ..., K$ we find the empirical risk minimizer $\widehat{h}_k = \text{argmin}_{h \in H_k} \text{error}_{\text{train}}(h)$ within H_k. Then we find

$$\widehat{k} = \underset{k=1,...,K}{\text{argmin}} \left(\text{error}_{\text{train}}(\widehat{h}_k) + \varepsilon(H_k, m, \delta_k) \right),$$

(for some $\delta_1, ..., \delta_K$), and the structural risk minimizer is $\widehat{h} = \widehat{h}_{\widehat{k}}$.

Another possible procedure would be to simply use the empirical risk minimizer in the union of $H_1, ..., H_K$. In the next two problems, we'll try to see why structural risk minimization might be a better idea.

(a) **[5 points]** Let

$$\widehat{h}_{\text{union}} = \operatorname*{argmin}_{h \in \bigcup_{k=1}^{K} H_k} \text{error}_{\text{train}}(h)$$

and

$$h^*_{\text{union}} = \operatorname*{argmin}_{h \in \bigcup_{k=1}^{K} H_k} \text{error}_{\text{true}}(h).$$

Show an upper bound on $\text{error}_{\text{true}}(\widehat{h}_{\text{union}})$ in terms of $\text{error}_{\text{true}}(h^*_{\text{union}})$ (and some other terms) that holds with probability $\geq 1 - \delta$ for given δ. How does your bound simplify in the case that $H_1 \subseteq H_2 \subseteq ... \subseteq H_K$?

Let H_k be the set of "interval classifiers" (as defined in Problem 2.1(b)) with *up to* k intervals, and let $K = 100$. Give a lower bound on the number of samples m needed that is sufficient to *guarantee* that $\text{error}_{\text{true}}(\widehat{h}_{\text{union}}) - \text{error}_{\text{true}}(h^*_{\text{union}}) \leq 0.25$ with probability at least 0.95. **Answer:** Applying the above bound with $H = \bigcup_{k=1}^{K} H_k$,

$$\text{error}_{\text{true}}(\widehat{h}_{\text{union}}) \leq \text{error}_{\text{true}}(h^*_{\text{union}}) + 2\varepsilon\left(\bigcup_{k=1}^{K} H_k, m, \delta\right)$$

with probability $\geq 1 - \delta$. If $\bigcup_{k=1}^{K} H_k = H_K$, this simplifies to

$$\text{error}_{\text{true}}(\widehat{h}_{\text{union}}) \leq \text{error}_{\text{true}}(h^*_{\text{union}}) + 2\varepsilon\left(H_K, m, \delta\right).$$

For the interval classifiers with $K = 100$ and $\delta = 0.05$, we need m such that

$$16\sqrt{\frac{200\left(\ln\frac{m}{200} + 1\right) + \ln\frac{8}{0.05}}{2m}} \leq 0.25.$$

We can solve this numerically to see that $m \geq 4527173$ is sufficient.

(b) **[5 points]** Give a lower bound on the number of samples m so that the structural risk minimizer $\widehat{h}_{\widehat{k}}$, computed using $\delta_1 = ... = \delta_K = 0.05/100$ on the sequence of hypothesis spaces defined in part (a), satisfies $\text{error}_{\text{true}}(\widehat{h}_{\widehat{k}}) - \text{error}_{\text{true}}(h^*_{\text{union}}) \leq 0.25$ with probability at least 0.95, assuming that $h^*_{\text{union}} \in H_5$.

Answer: We have that with probability $\geq 1 - \sum_{k=1}^{K} \delta_k$,

$$\begin{aligned}\text{error}_{\text{true}}(\widehat{h}_{\widehat{k}}) &\leq \min_k \left\{ \min_{h \in H_k} \text{error}_{\text{true}}(h) + 2\varepsilon(H_k, m, \delta_k) \right\} \\ &\leq \min_{h \in H_{k'}} \text{error}_{\text{true}}(h) + 2\varepsilon(H_{k'}, m, \delta_{k'})\end{aligned}$$

for any fixed k'. In particular, the inequality holds for $k' = 5$. Also, since $h^*_{\text{union}} \in H_5$, $\text{error}_{\text{true}}(h^*_{\text{union}}) = \min_{h \in H_{k'}} \text{error}_{\text{true}}(h)$, so

$$\begin{aligned}\text{error}_{\text{true}}(\widehat{h}_{\widehat{k}}) &\leq \text{error}_{\text{true}}(h^*_{\text{union}}) + 2\varepsilon(H_5, m, \delta_5) \\ &= \text{error}_{\text{true}}(h^*_{\text{union}}) + 16\sqrt{\frac{10\left(\ln \frac{m}{10} + 1\right) + \ln \frac{8}{0.05/100}}{2m}}\end{aligned}$$

Solving numerically, we see that $m \geq 247493$ suffices, which is an order of magnitude fewer samples than was required without structural risk minimization in part (a).

3 Boosting [25pt, Derry]

Consider a stepwise algorithm **A**:

Input parameters: T, $\mathcal{H}$, ϕ

Initialize the classifier $f_0(x) = 0$

for $t = 1$ to T **do**:

1. Compute

$$(h_t, \alpha_t) = argmin_{\alpha \in \mathbb{R}, h \in \mathcal{H}} \sum_{i=1}^{m} \phi(y_i, f_{t-1}(x_i) + \alpha h(x_i))$$

2. Update the classifier

$$f_t(x) = f_{t-1}(x) + \alpha_t h_t(x)$$

end for

return the classifier $sign(f_T(x))$

The intuition is that, at each step, the algorithm greedily adds a hypothesis $h \in \mathcal{H}$ to the current hypothesis to minimize the ϕ-risk.

1. **[5 pts]** What would be the form of $\phi(y, y')$ that will make algorithm **A** equivalent to AdaBoost?

   Answer: Exponential loss, $\phi(y, y') = exp(-y'y)$

2. **[10 pts]** Using the risk function you have defined above, prove that AdaBoost is equivalent to algorithm **A**.

   Hint: Work out the value of h_t that will minimize the risk function $\phi(y, y')$ for any fixed value of $\alpha > 0$ (further hints: h_t is not a function on α). Then, given this h_t find the α_t that will minimize the risk function $\phi(y, y')$. Think also how the weights $D_t(i)$ in AdaBoost is related to algorithm **A**.

   Answer:

$$
\begin{aligned}
(h_t, \alpha_t) &= argmin_{\alpha \in \mathbb{R}, h \in \mathcal{H}} \sum_{i=1}^{m} \phi(y_i, f_{t-1}(x_i) + \alpha h(x_i)) \\
&= argmin_{\alpha \in \mathbb{R}, h \in \mathcal{H}} \sum_{i=1}^{m} exp(-y_i(f_{t-1}(x_i) + \alpha h(x_i))) \\
&= argmin_{\alpha \in \mathbb{R}, h \in \mathcal{H}} \sum_{i=1}^{m} exp(-y_i f_{t-1}(x_i)) exp(-y_i \alpha h(x_i))) \\
&= argmin_{\alpha \in \mathbb{R}, h \in \mathcal{H}} \sum_{i=1}^{m} D_{t-1}(i) exp(\alpha) \mathbb{1}\{y_i \neq h(x_i)\} + \sum_{i=1}^{m} D_{t-1}(i) exp(-\alpha)\{1 - \mathbb{1}\{y_i \neq h(x_i)\}\} \\
&= argmin_{\alpha \in \mathbb{R}, h \in \mathcal{H}} exp(-\alpha) \sum_{i=1}^{m} D_{t-1}(i) + (exp(\alpha) - exp(-\alpha)) \sum_{i=1}^{m} D_{t-1}(i) \mathbb{1}\{y_i \neq h(x_i)\}
\end{aligned}
$$

where

$$
D_{t-1}(i) = \frac{exp(-y_i(f_{t-1}(x_i))}{\sum_{i=1}^{m} exp(-y_i(f_{t-1}(x_i))}
$$

$$
h(x_i) \in \{-1, +1\}
$$

Hence for any fixed value of $\alpha > 0$,

$$
\begin{aligned}
h_t &= argmin_{h \in \mathcal{H}} (exp(\alpha) - exp(-\alpha)) \sum_{i=1}^{m} D_{t-1}(i) \mathbb{1}\{y_i \neq h(x_i)\} \\
&= argmin_{h \in \mathcal{H}} \sum_{i=1}^{m} D_{t-1}(i) \mathbb{1}\{y_i \neq h(x_i)\}
\end{aligned}
$$

Given h_t, to obtain α_t,

$$
\begin{aligned}
\alpha_t &= argmin_{\alpha \in \mathbb{R}} exp(\alpha) \sum_{i=1}^{m} D_{t-1}(i) \mathbb{1}\{y_i \neq h_t(x_i)\} + exp(-\alpha) \sum_{i=1}^{m} D_{t-1}(i)\{1 - \mathbb{1}\{y_i \neq h_t(x_i)\}\} \\
&= argmin_{\alpha \in \mathbb{R}} exp(\alpha)\epsilon_t + exp(-\alpha)(1 - \epsilon_t)
\end{aligned}
$$

where

$$
\epsilon_t = \sum_{i=1}^{m} D_{t-1}(i) \mathbb{1}\{y_i \neq h_t(x_i)\}
$$

To find the minimum, differentiating the expression $exp(\alpha)\epsilon_t + exp(-\alpha)(1-\epsilon_t)$ w.r.t. α and setting it to zero,

$$\begin{aligned} exp(\alpha)\epsilon_t - exp(-\alpha)(1-\epsilon_t) &= 0 \\ exp\,(2\alpha) &= \frac{1-\epsilon_t}{\epsilon_t} \\ \alpha_t &= \frac{1}{2}\,log\,\frac{1-\epsilon_t}{\epsilon_t} \end{aligned}$$

which is similar to AdaBoost.

3. **[10 pts]** Now consider a more general algorithm where $h_t \in \mathcal{H}$ from $t = 1$ to T be any arbitrary sequence of classifiers. Let $\{x_i, y_i\}_{i=1}^m$ be a training set of m observations. Starting with $f_0 = 0$, f_t is recursively defined as $f_t = \sum_{i=1}^t \alpha_i h_i$ and $\alpha_t = \beta\, log\, \frac{1-\epsilon_t}{\epsilon_t}$ where

$$\epsilon_t = \sum_{i=1}^m D_{t-1}(i)\mathbf{1}\{y_i \neq h_t(x_i)\}$$

which is the weighted training error of the classifier h_t. Prove that for all T:

$$\sum_{i=1}^m \frac{1}{m}\, exp\,(-\frac{1}{\beta} y_i f_T(x_i)) = 1$$

which implies that any sequence of classifiers can be combined linearly to form a good combination while maintaining a constant exponential loss on the data.

Answer:

Define:

$$\begin{aligned} D_0(i) &= \frac{1}{m} \\ D_1(i) &= \frac{\frac{1}{m}\, e^{-\alpha_1 \frac{1}{\beta} y_i h_1(x_i)}}{Z_1} \\ D_2(i) &= \frac{\frac{1}{m}\, e^{-\alpha_1 \frac{1}{\beta} y_i h_1(x_i)}\, e^{-\alpha_2 \frac{1}{\beta} y_i h_2(x_i)}}{Z_1 Z_2} \\ &\ldots \\ D_T(i) &= \frac{\frac{1}{m}\, e^{-\frac{1}{\beta} y_i \sum_{t=1}^T \alpha_t h_t(x_i)}}{\prod_{t=1}^T Z_t} \end{aligned}$$

where,

$$Z_t = \sum_{i=1}^m D_{t-1}(i) e^{-\alpha_t \frac{1}{\beta} y_i h_t(x_i)}$$

Since $\sum_{i=1}^{m} D_T(i) = 1$,

$$\sum_{i=1}^{m} \frac{1}{m} e^{-\frac{1}{\beta} y_i \sum_{t=1}^{T} \alpha_t h_t(x_i)} = \prod_{t=1}^{T} Z_t$$

Now,

$$\begin{aligned}
Z_t &= \sum_{i=1}^{m} D_{t-1}(i) e^{-\alpha_t \frac{1}{\beta} y_i h_t(x_i)} \\
&= \sum_{i=1}^{m} D_{t-1}(i) e^{\alpha_t \frac{1}{\beta}} \mathbb{1}\{y_i \neq h_t(x_i)\} + \sum_{i=1}^{m} D_{t-1}(i) e^{-\alpha_t \frac{1}{\beta}} \{1 - \mathbb{1}\{y_i \neq h_t(x_i)\}\} \\
&= e^{\alpha_t \frac{1}{\beta}} \sum_{i=1}^{m} D_{t-1}(i) \mathbb{1}\{y_i \neq h_t(x_i)\} + e^{-\alpha_t \frac{1}{\beta}} \sum_{i=1}^{m} D_{t-1}(i) \{1 - \mathbb{1}\{y_i \neq h_t(x_i)\}\} \\
&= \frac{1-\epsilon_t}{\epsilon_t} \sum_{i=1}^{m} D_{t-1}(i) \mathbb{1}\{y_i \neq h_t(x_i)\} + \frac{\epsilon_t}{1-\epsilon_t} \sum_{i=1}^{m} D_{t-1}(i) \{1 - \mathbb{1}\{y_i \neq h_t(x_i)\}\} \quad (\text{since } \alpha_t = \beta \, log \, \frac{1-\epsilon_t}{\epsilon_t}) \\
&= \frac{1-\epsilon_t}{\epsilon_t} \epsilon_t + \frac{\epsilon_t}{1-\epsilon_t} (1-\epsilon_t) \quad (\text{since } \epsilon_t = \sum_{i=1}^{m} D_{t-1}(i) \mathbb{1}\{y_i \neq h_t(x_i)\} \text{ and } \sum_{i=1}^{m} D_{t-1}(i) = 1) \\
&= 1 - \epsilon_t + \epsilon_t \\
&= 1
\end{aligned}$$

Hence,

$$\sum_{i=1}^{m} \frac{1}{m} exp\left(-\frac{1}{\beta} y_i f_T(x_i)\right) = \prod_{t=1}^{T} Z_t = \prod_{t=1}^{T} 1 = 1$$

4 PCA [Avi 25 pts]

In this question we will try to understand PCA by showing two cool ways of interpreting the first principal component. One is the direction of maximum variance after projection and the second is the direction that minimizes reconstruction error. Note that the first principal component is the first eigenvector of the sample covariance matrix.

Consider n points $X_1, ..., X_n$ in p-dimensional space, and let X be the $n \times p$ matrix representing these points. Assume that the data points are centered, ie, $\vec{1}^\top X = \vec{0}$. Consider a unit vector $v \in \mathbb{R}^p$ and project all the points onto this vector (hence every point becomes a one-dimensional point on the direction of unit vector v).

1 [1 pt] Argue that the projection is given by Xv.

Soln:- Let us decompose the vector representing X_i into two orthogonal vectors X_{iv} and $X_{iv'}$ where X_{iv} is parallel to v. Using notation that $X = [X_1^T, X_2^T, \ldots, X_n^T] = [X_i^T]$ we get

$$Xv = [(X_{iv} + X_{iv'})]v = [X_{iv}^T]v + [X_{iv'}^T]v = [X_{iv}^T]v = X_v v$$

Given that v is a unit vector, $X_v v$ given the component of X in direction v. Since $Xv = X_v v$, Xv represents projection of X onto v.

2 [2 pt] What is the sample mean of all the points after the projection?

Soln:- Sample mean after projection is given by

$$\frac{1}{n}[\overrightarrow{1}^T(Xv)] = \frac{1}{n}[(\overrightarrow{1}^T X)v] = \frac{1}{n}[\overrightarrow{0}] = 0$$

3 [2 pt] What is the sample variance of all the points after the projection?

Soln:- Given that the sample mean is zero we can write the variance as

$$\frac{1}{n}[(Xv)^T(Xv)] = \frac{1}{n}[v^T X^T X v] = v^T \left[\frac{X^T X}{n}\right] v = v^T \Sigma v$$

where $\Sigma = X^T X$ is the sample variance of original p-dimensional points (X).

4 [2 pt] Setup the problem of maximizing the sample variance of the projection onto v subject to a constraint on the L2-norm of v.

Soln:-

$$\begin{aligned} &\max_v v^T \Sigma v \\ &st. \|v\|^2 = 1 \end{aligned} \tag{1}$$

5 [4 pt] Solve the minimization problem to show that the solution is the first PC. (Hint: take the Lagrangian of the above problem, differentiate and substitute to zero, to get to the optimum solution)

Soln:- By stationarity, at optimality we have

$$2\Sigma v^* + \lambda^* v^* = 0$$

Thus the optimal value is $v^{*T}\Sigma v^* = \lambda$ and so the vector that maximizes variance after projection, is the eigenvector associated with the largest eigenvalue λ of the covariance matrix Σ.

So we have now proved that the direction of maximum covariance is the first PC. Now we show that the direction that minimizes reconstruction error is also the first PC.

6 [1 pt] Argue that the reconstruction of X_i using v is $(X_i^\top v)v$.

Soln:- The reconstruction error of X_i using v can be written as the following optimization problem (with α being scalar): $\min_\alpha \|X_i - \alpha v\|^2$. Taking derivative wrt α and setting it to zero gives us the following:

$$2(X_i - \alpha v)^T v = 0 \Leftrightarrow X_i^T v = \alpha v^T v \Leftrightarrow \alpha = X_i^T v$$

Since v is a unit vector $v^T v = 1$. So, $\alpha v = (X_i^T v)v$ is the reconstruction of X_i using v.

7 [2 pt] You projected X_i to $X_i^\top v$ and then reconstructed it using $(X_i^\top v)v$. What is the reconstruction error of X_i, when measured in L2-norm?

Soln:-

$$\|(X_i^T v)v - X_i\|_2$$

8 [2 pt] What is the total squared reconstruction error over all points? **Soln:-**

$$\|(X^T v)v - X\|_F^2$$

9 [2 pt] Show that minimizing total squared reconstruction error is equivalent to minimizing $-\|Xv\|_2^2$.

Soln:-

$$\begin{aligned}
\|(X^T v)v - X\|_F^2 &= tr(((X^T v)v - X)^T((X^T v)v - X)) \\
&= tr(vv^T X^T Xvv^T) - 2tr(vv^T X^T X) + tr(X^T X) \\
&= tr(v^T X^T Xvv^T v) - 2tr(v^T X^T Xv) + tr(X^T X) \\
&= tr(v^T X^T Xv) - 2tr(v^T X^T Xv) + tr(X^T X) \\
&= -tr(v^T X^T Xv) + tr(X^T X) \\
&= -\|Xv\|_2^2 + \|X\|_2^2
\end{aligned}$$

since the minimization is wrt to v, $\|X\|_2^2$ is constant.

10 [4 pt] Solve the minimization problem to show that the solution is the first PC. (Hint: take the Lagrangian of the above problem, differentiate and substitute to zero, to get to the optimum solution)

Soln:- The optimization problem is the same as in part 5.

4.1 [3 pts] SVD and PCA

Let us define a new variable Y as

$$Y = X^T \tag{2}$$

where X is a $n \times p$ matrix containing the data points as defined before. If the SVD of Y is given by $Y = U\Sigma V^T$ then show that the columns of V are the PCA of X.

Soln:

$$XX^T = Y^T Y = V\Sigma UU^T \Sigma V^T = V\Sigma^2 V^T \tag{3}$$

Thus $(XX^T)V_i = \Sigma_{ii}^2 V_i$ which implies that the columns of V are the PCA of X.