CS771: Machine learning: tools, techniques and applications

Mid-semester exam

Time: 2 hours 15-Feb-2015
Max marks: 80

1. *Answer all 4 questions. The question paper has 2 pages.*
2. *Your answers should show all calculations for full credit.*
3. *Please be precise in your answers.*
4. *You can consult* **only handwritten class notes**. *Any other printed or digital material or electronic gadgets are not allowed.*

1. (a) If $|\mathcal{L}| = n$ then answer the following with respect to the bagged learning sets used for creating the decision trees of a random forest classifier:
   i. What is the expression for the average size of the bagged learning set?
   ii. What is the minimum size of a bagged learning set?
   iii. What is the maximum size of a bagged learning set?

Solution:
i) Average size of bagged: $n(1 - \frac{1}{e})$.
ii) Min. size of bagged set: 1
iii) Max. sized of bagged set: n

(b) What is the *horizon effect* and how does it affect the construction of decision trees (DTs)? Assume a DT has been grown fully. Assuming a validation set is available how will you use it to decide whether or not to prune a sub-tree subtended from a node z in the DT?

Solution:
The horizon effect is a situation where the maximum impurity reduction at a particular level or depth is less than the maximum impurity reduction at a deeper/higher depth level. This normally means that a threshold based stopping criterion does not give the best decision tree. To avoid the horizon effect we grow the tree fully and then prune.

Find the error rate for vectors from the validation set that end up at node z and are classified at that node (i.e. tree subtended from z has been pruned) - let this be e_{pruned}. Also find the error rate when the classification is done at the leaves of the tree subtended from z - let this error rate be $e_{unpruned}$. If $e_{pruned} \leq e_{unpruned}$ prune the tree else do not.

(c) Consider the function $\phi(x) = max(x, 1-x),\ 0 \leq x \leq 1$. Can ϕ be an impurity function? Justify your answer.

Let a node z in a DT contain 10 vectors of class ω_1 and 10 vectors of class ω_2, written as $(10, 10)$ where the first number gives the number of vectors of class ω_1 and the second one the number of vectors of class ω_2 at a node. While evaluating possible binary splits for z we get the following splits: i) $(3, 5)(7, 5)$ ii) $(4, 3)(6, 7)$ iii) $(2, 5)(8, 5)$. If Gini impurity is being used which of the 3 splits will you choose? Justify.

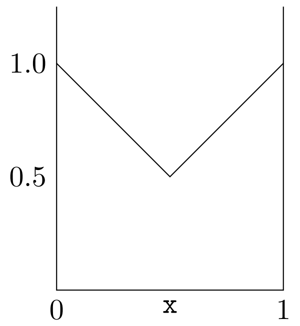


Figure 1: Graph of $\phi = max(x, 1 - x)$, $0 \leq x \leq 1$

.

Solution:

The function ϕ is clearly defined for binary classification with x being the fraction of vectors of class ω_1, and $(1 - x)$ the fraction of vectors of class ω_2.

The graph of the function is shown above. It does not satisfy two important conditions for an impurity function. First impurity should be maximum when the two classes are equally represented that is at $x = 0.5$. Actually, it is a minimum at that point and second when nodes are pure, that is at $x = 0.0$ and $x = 1.0$, impurity should be a minimum while it is a maximum. So ϕ cannot be an impurity function. Actually, it is a mirror image of an impurity function w.r.t to the X-axis.

To calculate the correct split we do not actually need to go through the impurity calculations. We just have to pick the minimum impurity split. This will result in the maximum impurity reduction. By inspection it is clear that the impurity will be smallest when the disbalance between the vectors of class ω_1 and ω_2 is largest at the child nodes (nodes are more pure when disbalance between classes is higher). The disbalance is a maximum for split iii) and so this will have the minimum impurity (giving rise to maximum impurity reduction) and so iii) should be the chosen split.

Of course, one can confirm this with a calculation:

Impurity for split i): $\frac{30}{64} + \frac{70}{144}$
Impurity for split ii): $\frac{24}{49} + \frac{84}{169}$
Impurity for split iii): $\frac{20}{49} + \frac{80}{169}$

It is clear that iii) is the minimum impurity value.

(d) We had an upper bound $PE^* \leq \frac{\overline{\rho}(1-s^2)}{s^2}$ for the generalization error of a random forest, where $\overline{\rho}$ is the average correlation between pairs of trees in the forest and s is the average strength of a tree in the forest. Given the learning set $\mathcal{L}$ briefly describe how you will estimate the error bound above.

Solution:

The estimate of the bound can be calculated by estimating the values of s and $\overline{\rho}$ by using their definitions. Let $|\mathcal{L}| = n$. First we estimate s. $s = E_{\mathbf{X},Y}[mr(\mathbf{X}, Y)]$ where $mr(\mathbf{X}, Y) = P_\theta(f(\mathbf{X}, \theta) = Y) - \underset{j \neq Y}{max} P_\theta(f(\mathbf{X}, \theta) = j)$. Let estimate of $P_\theta(f(\mathbf{X}, \theta) = Y)$ be written as $\hat{p}(\mathbf{x}, y)$.

Let $\mathcal{L}_i$ be the i^{th} bagged sample and $f(\mathbf{x}, \theta_i)$ the corresponding DT. Then we can calculate

$\hat{p}(\mathbf{x}, y)$ using the OOB vector $(\mathbf{x}, y)$:

$$\hat{p}(\mathbf{x}, y) = \frac{\sum_i I_{[f(\mathbf{x},\theta_i)=y;(\mathbf{x},y)\notin\mathcal{L}_i]}}{\sum_i I_{[(\mathbf{x},y)\notin\mathcal{L}_i]}}$$

Then the strength s can be estimated using the learning set $\mathcal{L}$:

$$\hat{s} = \frac{1}{n}\sum_{i=1}^{n}(\hat{p}(\mathbf{x}_i, y_i) - \hat{p}(\mathbf{x}_i, \tilde{y_i})) \quad \text{where} \quad \tilde{y_i} = \underset{y_i' \neq y_i}{argmax}\, \hat{p}(\mathbf{x}_i, y_i')$$

To estimate $\bar{\rho}$ we use $\bar{\rho} = \frac{Var(mr(\mathbf{X},Y))}{E_\theta[\sigma(\theta)]^2}$. We have:

$$\begin{aligned} Var(mr(\mathbf{X}, Y)) &= E_{\mathbf{X},Y}[mr(\mathbf{X}, Y)^2] - s^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(\hat{p}(\mathbf{x}_i, y_i) - \hat{p}(\mathbf{x}_i, \tilde{y_i}))^2 - \hat{s}^2 \end{aligned}$$

To calculate the denominator $E_\theta[\sigma(\theta)]^2$ we use the expression for $\sigma(\theta)$:

$$\begin{aligned} \sigma(\theta)^2 &= Var_{\mathbf{X},Y}(rmg(\mathbf{X}, Y, \theta)) \\ &= E_{\mathbf{X},Y}[rmg(\mathbf{X}, Y, \theta)^2] - (E_{\mathbf{X},Y}[rmg(\mathbf{X}, Y, \theta)])^2 \end{aligned}$$

Note that the raw margin score is defined as: $rmg(\mathbf{X}, Y, \theta) = I(f(\mathbf{X}, \theta) = Y) - I(f(\mathbf{X}, \theta) = \tilde{Y})$ where $\tilde{Y}$ is the label different from Y that has the maximum probability of being predicted. The two terms of rmg can be estimated by using OOB vectors for the ith bagged set $\mathcal{L}_i$ to get I_Y and $I_{\tilde{Y}}$. The expectation $E_{\mathbf{X},Y}[.]$ can be calculated by dividing I_Y ($I_{\tilde{Y}}$) by n_i the number of OOB vectors in $\mathcal{L}_i$. Finally, to get the expectation $E_\theta[.]$ we find the average w.r.t all the θs - that is all the DTs in the forest.

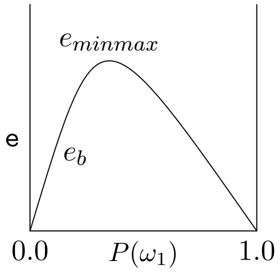
[4,(2+3),(2+3),6=20]

2. BDR requires knowing class conditional PDFs and a priori probabilities. In some cases the a priori probabilities may not be known and it may not be possible to calculate reasonable estimates from what is known. In such cases one can choose to minimize the maximum error (or more generally loss) that can be incurred due to misclassification. In a 2-class case if the decision regions R_1 and R_2 are known or given then we know the error e can be written as:

$$e = P(\omega_1)\int_{R_2} p(\mathbf{x}|\omega_1)d\mathbf{x} + P(\omega_2)\int_{R_1} p(\mathbf{x}|\omega_2)d\mathbf{x} = P(\omega_1)r_{12} + P(\omega_2)r_{21}$$

where $P(\omega_2) = 1 - P(\omega_1)$. If R_1 and R_2 are given/known then e is a linear function of $P(\omega_1)$ and is a maximum either at $P(\omega_1) = 0$ or $P(\omega_1) = 1$. The Bayes error, e_b is bound to be less than or equal to e for any value of $P(\omega_1)$. If we choose the regions R_1, R_2 or equivalently $P(\omega_1)$ such that the Bayes error is a maximum (denoted by e_{minmax}) then the actual error for any value of $P(\omega_1)$ will always be less than equal to e_{minmax} and it is independent of the value of the priors. See the figure below.

(a) Formulate the binary minmax decision problem for the setting where the loss coefficients are given by $\lambda_{ij},\ i, j = 1..2$ and derive the expression for the minmax risk and the expression to calculate the decsion boundary.



Solution:

Let R_1 be the region where the classifier predicts ω_1 and R_2 where it predicts ω_2. Then the total risk or loss r with loss coefficients $\lambda_{ij},\ i,j = 1..2$ is:

$$r = \int_{R_1} (\lambda_{11}P(\omega_1)p(\mathbf{x}|\omega_1) + \lambda_{21}P(\omega_2)p(\mathbf{x}|\omega_2))d\mathbf{x} + \int_{R_2} (\lambda_{12}P(\omega_1)p(\mathbf{x}|\omega_1) + \lambda_{22}P(\omega_2)p(\mathbf{x}|\omega_2))d\mathbf{x}$$

Now use the fact that $P(\omega_2) = 1 - P(\omega_1)$, $\int_{R_1} p(\mathbf{x}|\omega_1)d\mathbf{x} + \int_{R_2} p(\mathbf{x}|\omega_1)d\mathbf{x} = 1$ and similarly for $p(\mathbf{x}|\omega_2)$ and separate terms containing $P(\omega_1)$ and those not containing $P(\omega_1)$:

$$\begin{aligned} r =& \lambda_{22} + (\lambda_{21} - \lambda_{22}) \int_{R_1} p(\mathbf{x}|\omega_2)d\mathbf{x} + \\ & P(\omega_1)\left[\lambda_{11} \int_{R_1} p(\mathbf{x}|\omega_1)d\mathbf{x} - \lambda_{21} \int_{R_1} p(\mathbf{x}|\omega_2)d\mathbf{x} + \lambda_{12} \int_{R_2} p(\mathbf{x}|\omega_1)d\mathbf{x} - \lambda_{22} \int_{R_2} p(\mathbf{x}|\omega_2)d\mathbf{x}\right] \end{aligned}$$

Notice that the risk is a linear function of $P(\omega_1)$ assuming $R_1,\ R_2$ are given.
We want the minmax risk r_{minmax} to be independent of $P(\omega_1)$ so the minmax risk is given by the first term that does not contain $P(\omega_1)$. And for the minmax solution since the risk must be independent of $P(\omega_1)$ the second term within [..] that multiplies $P(\omega_1)$ should be 0 and this gives the regions $R_1,\ R_2$.
So,

$$r_{minmax} = \lambda_{22} + (\lambda_{21} - \lambda_{22}) \int_{R_1} p(\mathbf{x}|\omega_2)d\mathbf{x}$$

and the regions R_1 and R_2 are given by:

$$\lambda_{11} \int_{R_1} p(\mathbf{x}|\omega_1)d\mathbf{x} - \lambda_{21} \int_{R_1} p(\mathbf{x}|\omega_2)d\mathbf{x} + \lambda_{12} \int_{R_2} p(\mathbf{x}|\omega_1)d\mathbf{x} - \lambda_{22} \int_{R_2} p(\mathbf{x}|\omega_2)d\mathbf{x} = 0$$

The above equation can be slightly simplified by using the identities for $p(\mathbf{x}|\omega_1)$ and $p(\mathbf{x}|\omega_2)$.

(b) If $p(\mathbf{x}|\omega_1) = \mathcal{N}(5, 1)$ and $p(\mathbf{x}|\omega_2) = \mathcal{N}(6, 1)$ what will be x_t (the boundary point) for the minmax criterion?

Solution:

For $0-1$ loss the equation for the decision regions reduces to:

$$\int_{R_1} p(\mathbf{x}|\omega_2)d\mathbf{x} = \int_{R_2} p(\mathbf{x}|\omega_1)d\mathbf{x}$$

When the class conditional distributions are univariate normal distributions with the same variance and different means μ_1 and μ_2 the integrals will be equal exactly at the point of intersection of the two distributions which by symmetry will be at the half way point between the means that is: $\frac{\mu_1+\mu_2}{2}$. For the given distributions for ω_1 and ω_2 this is: $x_t = \frac{5+6}{2} = 5.5$. For arbitrary distributions calculating the region boundary is non-trivial and closed form solutions are generally not possible.

[12,8=20]

3. For a 2-class problem the prior probabilities are: $P(\omega_1) = \frac{1}{4}$ and $P(\omega_2) = \frac{3}{4}$. The class conditional distributions for $\mathbf{x} = x$, that is $\mathbf{x}$ has only a single attribute, are:

$$p(x|\omega_1) = \mathcal{N}(0,1) \text{ and } p(x|\omega_2) = \mathcal{N}(1,1)$$

.

(a) Calculate the threshold boundary value x_t which gives the probability of minimum error.

Solution:
The decision boundary is given by: $P(\omega_1)p(\mathbf{x}|\omega_1) = P(\omega_2)p(\mathbf{x}|\omega_2)$.
Using the given normal distribution for ω_1 and ω_2 this gives us:

$$\begin{aligned} \frac{1}{4}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} =& \frac{3}{4}\frac{1}{\sqrt{2\pi}}e^{-\frac{(x-1)^2}{2}} \\ 1 =& 3\,e^{\frac{2x-1}{2}}. \quad \text{Taking ln of both sides.} \\ 0 =& ln(3) + \frac{2x-1}{2}. \quad \text{So, boundary } x_t \text{ is:} \\ x_t =& \frac{1}{2} - ln(3) \end{aligned}$$

(b) If the loss matrix is: $\lambda_{ij} = \begin{bmatrix} 0 & 1 \\ \frac{1}{2} & 0 \end{bmatrix}$ find the threshold boundary value x_t for minimum risk.

Solution:
In this case we get:

$$\begin{aligned} r_1 =& \lambda_{11}p(\mathbf{x}|\omega_1)P(\omega_1) + \lambda_{21}p(\mathbf{x}|\omega_2)P(\omega_2) \\ r_2 =& \lambda_{12}p(\mathbf{x}|\omega_1)P(\omega_1) + \lambda_{22}p(\mathbf{x}|\omega_2)P(\omega_2) \end{aligned}$$

The rule for prediction is: ω_1 if $r_1 < r_2$ and ω_2 if $r_2 < r_1$. The boundary point is given by solving $r_1 = r_2$. Using the given values for λ_{ij} and the class conditional distributions we get:

$$\begin{aligned}
\frac{1}{2}\frac{3}{4}\frac{1}{\sqrt{2\pi}}e^{-\frac{(\mathbf{x}-1)}{2}} &= \frac{1}{4}\frac{1}{\sqrt{2\pi}}e^{-\frac{\mathbf{x}^2}{2}} \\
\frac{3}{2}e^{\frac{2x-1}{2}} &= 1. \quad \text{Taking } ln \text{ of both sides.} \\
ln(\frac{3}{2}) + \frac{2x-1}{2} &= 0. \quad \text{This gives the following value for } x_t\text{:} \\
x_t &= \frac{1}{2} - ln(\frac{3}{2})
\end{aligned}$$

[10,10=20]

4. In the proof of the perceptron convergence theorem we reached the following inequality:

$$\|\mathbf{w}'(t+1) - \alpha\mathbf{w}'^*\|^2 \leq \|\mathbf{w}'(t) - \alpha\mathbf{w}'^*\|^2 + \rho(t)^2\beta^2 - 2\rho(t)\alpha|\gamma|$$

with

$$\beta = \max_{\mathbf{Y}\in\mathcal{P}(\mathcal{L}),\mathbf{Y}\neq\Phi} \|\sum_{\mathbf{x}\in\mathbf{Y}} \delta_x \mathbf{x}'\|$$

and

$$\gamma = \max_{\mathbf{Y}\in\mathcal{P}(\mathcal{L}),\mathbf{Y}\neq\Phi} \sum_{\mathbf{x}\in\mathbf{Y}} \delta_x \mathbf{w}'^{*T}\mathbf{x}'$$

where $\mathcal{P}(\mathcal{L})$ is the power set of $\mathcal{L}$.

(a) We wish to explore a proof of the case when ρ is fixed and does not depend on t. Choose a suitable value for α (using β, γ) and a bound on ρ and argue that the perceptron algorithm converges in finitely many iterations.

Solution:
Choose $\alpha = \frac{\beta^2}{|\gamma|}$ and substitute for α in the equation. We replace $\rho(t)$ by ρ since it is a constant and not dependent on t any more.

$$\begin{aligned}
\|\mathbf{w}'(t+1) - \alpha\mathbf{w}'^*\|^2 &\leq \|\mathbf{w}'(t) - \alpha\mathbf{w}'^*\|^2 + \rho^2\beta^2 - 2\rho\beta^2 \\
&\leq \|\mathbf{w}'(t) - \alpha\mathbf{w}'^*\|^2 + \beta^2(\rho^2 - 2\rho) \\
&\text{unfolding with respect to } t \text{ gives:} \\
&\leq \|\mathbf{w}'(0) - \alpha\mathbf{w}'^*\|^2 + \beta^2 t(\rho^2 - 2\rho)
\end{aligned}$$

If the rhs has to be reduced to ≤ 0 then $(\rho^2 - 2\rho)$ must be negative. In which case for a large enough t, say $\hat{t}$ the rhs will be ≤ 0 thereby proving that the iteration converges. This implies $\rho^2 < 2\rho$ and $\rho > 0$ giving the following bound $0 < \rho < 2$.

(b) Based on a) above give an expression for the number of iterations that will be needed for convergence.

Solution:
In part a) we require that rhs goes to 0 at $\hat{t}$ so $\|\mathbf{w}'(0) - \alpha\mathbf{w}'^*\|^2 = \beta^2\rho(2-\rho)\hat{t}$ giving us the following value for $\hat{t}$:

$$\hat{t} = \left\lceil \frac{\|\mathbf{w}'(0) - \alpha\mathbf{w}'^*\|^2}{\beta^2\rho(2-\rho)} \right\rceil \quad \text{where} \quad 0 < \rho < 2$$

[12,8=20]