Machine Learning 10-701 Final Exam May 9, 2013

Name:

Instructions

- Anything on paper is OK in arbitrary shape size, and quantity.
- Electronic devices are not acceptable. This includes iPods, iPads, Android tablets, Blackberries, Nokias, Windows phones, Microsoft Surface, laptops, Chromebooks, MP3 players, digital cameras, Google Glass, or anything else that requires electricity.[1]
- **If we see you use any such device, the exam is over for you. Your results will not count and you will have failed this exam automatically, and there might me more serious consequences too. No exceptions. Switch off your phones before the exam.**
- The exam duration is 3 hours. Choose wisely the questions to answer first, by starting with the ones you believe you can solve easily.
- None of the questions should require more than 1 page to answer. It's OK to be concise if you address the key points of the derivation (it should be human-readable, though).

Grade:

- Problem 1: /20
- Problem 2: /10
- Problem 3: /10
- Problem 4: /10
- Problem 5: /10
- Problem 6: /10
- Problem 7: /10
- Problem 8: /10
- Problem 9: /10
- Problem 10: /10
- Problem 11: /5
- Problem 12: /5

Total: /120

[1] Obvious exceptions for pacemakers and hearing aids.

1 Short Answer

1. Describe the difference between non-parametric and parametric techniques. Why would use one over another (list pros & cons)?

   Parametric techniques assume some finite parameterization of the underlying data's distribution. Non-parametric techniques make no such assumptions (only implicit smoothness assumptions about underlying distribution). Note that both techniques actually have parameters to their estimators (e.g. k in k-nn is a parameter to a non-parametric technique.) Parametric techniques will typically have a faster rate of convergence (require less data). However, the rates of convergences are valid only if the parametric-model assumptions hold true. Thus, although the rates of nonparametric techniques are slower, one may get better estimators when unsure of the source of data.

2. Describe the difference between supervised and unsupervised learning. List examples of each.

   Supervised learning involves datasets of pairs of covariates (input features) and responses (output features) (x, y) respectively. Unsupervised tasks work only over datasets of input features. Examples of supervised learning are classification and regression. Examples of unsupervised learning are clustering, density estimation, dimensionality reduction, and outlier detection. Note that unsupervised learning is not limited to clustering/similarities/latent-variables.

3. In general, why does more data help ML tasks?

   Most ML tasks are comprised of estimators (of parameters, and of risks, for example). Results like LLN, error rates, and other concentration of measure bounds tell us that as sample sizes increase, so too will the quality of our estimates.

4. Describe what an estimator's bias and variance is. Why are these two quantities important?

   Bias is $(\mathbb{E}[\hat{\theta}] - \theta)$; roughly, it measures how good of an approximation the parameter space we are working in is to the true parameter space. E.g. a linear regressor will have a lot of bias when regressing a quadratic function (note this doesn't explicitly say anything about over-fitting). Variance is (wait for it) the variance of the estimator: $\mathrm{Var}[\hat{\theta}]$; intuitively, if the variance is high then the model changes a lot depending on the sample that is drawn (i.e. it over-fits). The real reason why we care about these quantities is that the make up the MSE $(\mathbb{E}[(\hat{\theta} - \theta)^2])$ as MSE = $\mathrm{Bias}^2 + \mathrm{Variance}$. We wouldn't care about these quantities if they didn't elucidate an important risk.

5. What is regularization, and what is its purpose? Give an example of regularization.

   Regularization penalizes model complexity in order to prevent over-fitting. An example of this is ridge regression.

6. What is the difference between PCA and ICA?

   Listing multiple of the many differences in the ICA slides would have sufficed. For example, PCA works over orthogonal directions, of varying importance, and can be used for dimensionality reduction; ICA need not have orthogonal directions, all are equally important, and can be used for separating independent sources (no compression).

7. Qualitatively describe why it may be a good idea to maximize the margin in a classifier.

   In short, it helps over-fitting and generalization. In fact, assuming that data is coming for a KDE of the training data with bandwidths going to zero (a good assumption if working with a lot of data) maximizing the margin is the optimal linear classifier. Intuitively, it's focusing in on the hard cases.

8. What does the kernel trick allow for? How is it used in practice?

   Many estimators and optimization problems in ML can be written out in terms involving inner products of the data. Hence, if we write out the inner products of higher dimensional non-linear mappings of the data, then we can build richer models. The kernel trick allows use to compute these inner products without explicitly computing their corresponding feature mappings. For example, kernerlized SVMs, kernel PCA use the kernel trick. Note that the kernel trick is for more than just classification.

9. Describe the difference between MLE and MAP.

   MLE finds the parameter the maximizes the likelihood of the random sample: $\hat{\theta}_{MLE} = \operatorname{argmax}_{\hat{\theta}} P(X|\hat{\theta})$. MAP finds the parameter the maximizes the posterior of the (now also) random parameter: $\hat{\theta}_{MAP} = \operatorname{argmax}_{\hat{\theta}} P(\hat{\theta}|X) = \operatorname{argmax}_{\hat{\theta}} P(X|\hat{\theta})P(\hat{\theta})$. Note that MAP incorporates a prior $P(\hat{\theta})$.

10. Why do we like the VC dimension? Why do we want to know it and how is it used?

    VC dimension quantifies the complexity/expressiveness (of even infinite) hypothesis spaces. It can be used in PAC Bounds to analyze sample size effects on risk estimation.

2 Bias/Variance Decomposition

- Figure 1 shows two curves (solid lines) fitting the same set of data (circles, dotted line shows the true curve). Which curve has higher bias and which one has higher variance?

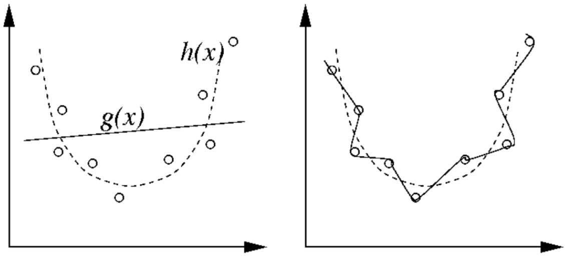


Figure 1: Bias/Variance Decomposition

- Suppose $X_1, ..., X_n \sim \mathcal{N}(\mu, \sigma^2)$.
  - Please come up with an unbiased estimate $\bar{X}$ of μ and explain why it is unbiased.
  - We usually use $S^2 = \frac{1}{n-1}\sum_i (X_i - \bar{X})^2$ as an estimate of σ^2. Show whether it is biased or not. (*Hint: it might take a few minutes to do the calculation*)
  - Why do we usually use $n-1$ instead of n?

Solution: 1. The left figure has higher bias, the right one has higher variance.
2. (1)

$$\bar{X} = \frac{1}{n}\sum_i X_i$$

It is unbiased since $E(\bar{X}) = \frac{1}{n}\sum_i E(X_i) = \mu$.

(2) It is unbiased. First notice $E(X_i^2) = Var(X_i) + (EX_i)^2 = \sigma^2 + \mu^2$, $Var(\bar{X}) = \frac{1}{n^2}\sum_i Var(X_i) = \frac{\sigma^2}{n}$, and $E(\bar{X}^2) = Var(\bar{X}) + (E\bar{X})^2 = \frac{\sigma^2}{n} + \mu^2$.

$$\begin{aligned}
E(S^2) &= \frac{1}{n-1}\sum_i E(X_i - \bar{X})^2 = \frac{1}{n-1}\sum_i E(X_i^2 - 2X_i\bar{X} + \bar{X}^2) \\
&= \frac{1}{n-1}\sum_i (\sigma^2 + \mu^2 - 2 * \frac{1}{n}(E(X_i^2) + \sum_{i \neq j} E(X_iX_j)) + \mu^2 + \frac{\sigma^2}{n}) \\
&= \frac{1}{n-1}\sum_i (\sigma^2 + \mu^2 - 2 * \frac{1}{n}(\sigma^2 + \mu^2 + (n-1)\mu^2) + \mu^2 + \frac{\sigma^2}{n}) \\
&= \frac{1}{n-1}\sum_i (\sigma^2 - \frac{1}{n}\sigma^2) = \sigma^2
\end{aligned}$$

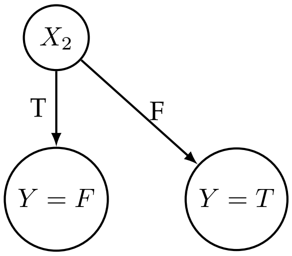
(3) We usually use $n-1$ since we want an unbiased estimator.

3 Decision Trees

Given the following observed outputs, build the simplest decision tree (*hint:* you can do this without explicitly computing information gain).

X_1	X_2	X_3	$Y = T$	$Y = F$
T	T	T	0	2
T	T	F	0	2
T	F	T	1	1
T	F	F	2	0
F	T	T	0	2
F	T	F	0	2
F	F	T	1	1
F	F	F	2	0

The simplest decision tree is:



4 K-means for Big Data

Given n samples $x_1, x_2, \ldots, x_n \in \mathbb{R}^p$, k-means tries to group these samples into k clusters. A widely used algorithm is the k-means algorithm. It consists of two stages: it first adjusts the cluster centers, then assigns all samples into the nearest clusters.

Now consider the case we have too many samples to fit into the memory, or even hard disk of a single machine. So we give you m machines. Please describe how to implement a distributed k-means algorithm, and answer the following problem

1. How do you partition the samples into m machines
2. What is the communication cost for one iteration of the k-means algorithm

Solution. Each machine get a part of samples, and its own version of cluster centers. On each iteration, all machines first calculate the new cluster centers based on their own data simultaneously, next a global consensus of these centers are obtained, then each machine reassign their samples.

To obtain a global consensus, a typically way is that, first each machine sends its cluster centers into a master machine, next the master averages these centers to get the global centers (weighted by #samples each machine has), then this master broadcasts the global centers.

It can be shown that this algorithm equals to the standard k-means algorithm.

Each machine needs to send out its cluster centers, and receive them after the global consensus is calculated. Therefore the total communication volume is $O(kmp)$.

5 Bayes-Ball

1. You are given the Bayesian Network in Figure 2.

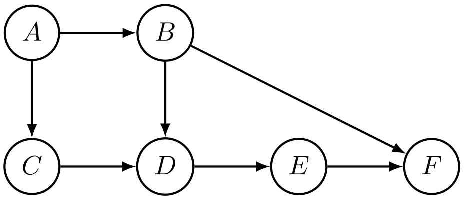


Figure 2: Bayesian model for independence question

Specify whether each of the following independence statements is true or false.

(a) $A \perp F$
Answer: False. Path: A-C-D-E-F, A-B-D-E-F

(b) $B \perp C|A$
Answer: True.

(c) $A \perp F|B, C$
Answer: True.

(d) $A \perp E|C$
Answer: False. Path: A-B-D-E.

(e) $A \perp E|C, F$
Answer: False. Path: A-B-D-E.

Each true/false question is worth 1.5 points.

2. Which of the structures in Figure 3 are equivalent? By equivalent we mean that they encode the same independence assumtions.

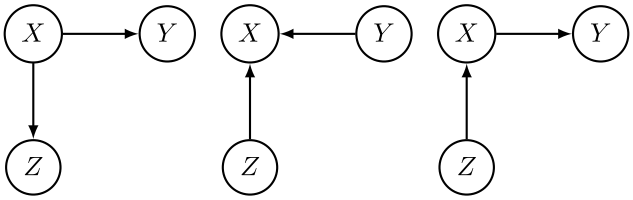


Figure 3: Bayesian models for equivalence question

Answer: The first and third models are equivalent. They encode the assumption $Y \perp Z|X$. For the graph in the middle, Y is independent of Z *unless* X is observed; this is known as a V-structure and was covered in the recitation. This question is worth 2.5 points.

6 More Bayes Nets

You are trying to learn a classification model for a dataset containing features $X_1 \dots X_5$ and class variable Y. You train classifiers C_1 and C_2 using each of the two bayesian models shown in Figure 4. You know that the model on the right is closer to the underlying data structure. How do you expect the performance of the two classifiers to change as the number of available samples increases? We expect a qualitative not quantitative answer.

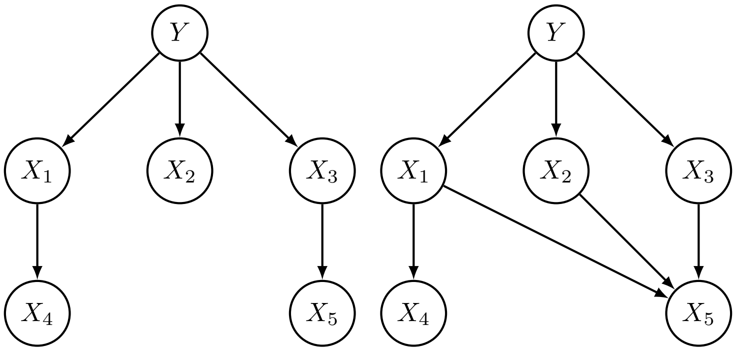


Figure 4: Bayesian models for classification model question

Answer: The model for C_1 (left) contains fewer parameters than C_2 (right). Given that C_2 is a more accurate representation, but also a more complex one, we can state the following:

- Both C_1 and C_2 will improve in accuracy as more data is available if they encode the correct independence assumptions about the data.
- At first (low sample size), C_1 will be more accurate because there are fewer parameters to estimate.
- As the samples size increases, C_2 will eventually outperform C_1.

Many students stated that one or both of the models 'overfit' with as more data becomes available. This is incorrect, as the structure is given, so the model cannot get any more complex than it is. If a Bayes Net model is 'wrong', its performance on training and testing data will be lower than that of a correct model. This is different from what is observed in the case of overfitting – low training error and high testing error.

7 Kernels

1. Let $k(x, y)$ be a kernel function, positive semidefinite function. Prove that

$$k(x, y)^2 \leq k(x, x)k(y, y) \quad \forall x, y$$

Answer:

$$k(x, y)^2 =< \phi(x), \phi(y) >^2 = ||\phi(x)||^2 ||\phi(y)||^2 (cos(\theta_{\phi(x),\phi(y)}))^2 \leq ||\phi(x)||^2 ||\phi(y)||^2 = k(x, x)k(y, y)$$

2. given a kernel $k(x, y)$ and a function $f(x)$, prove that $f(x)f(y)k(x, y)$ is also a kernel

Answer:

$$f(x)f(y)k(x, y) = f(x)f(y) < \phi(x), \phi(y) >=< f(x)\phi(x), f(y)\phi(y) >=< \phi'(x), \phi'(y) >$$

Therefore $f(x)f(y)k(x, y)$ is a kernel.

8 Maximum Likelihood Estimation

1. Let $X_1, X_2 \dots X_n \sim Uniform(-\theta, \theta)$ where $\theta > 0$.

   (a) Find the maximum likelihood estimator $\hat{\theta}_n$.
   Answer:
   $L(\theta|X) = \left(\frac{1}{2\theta}\right)^n I(\theta \geq T)$ where $T = \max\left\{|X_1|, \dots, |X_n|\right\}$.
   Thus, $\hat{\theta} = \max\left\{|X_1|, \dots, |X_n|\right\}$.

   (b) Show that the maximum likelihood estimator is consistent, i.e. $\lim_{n\to\infty} \hat{\theta}_n = \theta$.
   Answer:
   Let $Y_i = |X_i|$. $Y_i \sim$ Unif(0,θ).
   $P(|\hat{\theta} - \theta| > \epsilon) = P(|\max\left\{Y_1, \dots, Y_n\right\} - \theta| > \epsilon) = \prod_{i=1}^n P(Y_i < \theta - \epsilon) = \left(\frac{\theta-\epsilon}{\theta}\right)^n = \left(1 - \frac{\epsilon}{\theta}\right)^n \to 0$

2. Let $X \sim Bernoulli(\theta)$ be a single coin flip. Suppose that $\theta \in \Theta = \{1/3, 2/3\}$, meaning θ can take only 2 possible values.

   Find the maximum likelihood estimator $\hat{\theta}$.

   Answer:

$$L(\theta|X) = \theta^x (1-\theta)^{1-x} = \begin{cases} \left(\frac{1}{3}\right)^x \left(\frac{2}{3}\right)^{1-x} & \text{if } \theta = \frac{1}{3} \\ \left(\frac{2}{3}\right)^x \left(\frac{1}{3}\right)^{1-x} & \text{if } \theta = \frac{2}{3} \end{cases}$$

   For the given observation,

$$L(\theta|X=0) = \begin{cases} \frac{2}{3} & \text{if } \theta = \frac{1}{3} \\ \frac{1}{3} & \text{if } \theta = \frac{2}{3} \end{cases}$$

$$L(\theta|X=1) = \begin{cases} \frac{1}{3} & \text{if } \theta = \frac{1}{3} \\ \frac{2}{3} & \text{if } \theta = \frac{2}{3} \end{cases}$$

   Therefore,

$$\hat{\theta} = \begin{cases} \frac{1}{3} & \text{if } X = 0 \\ \frac{2}{3} & \text{if } X = 1 \end{cases}$$

9 Sampling

Our goal is to obtain samples from the 1-dimensional distribution having density $p(x)$ with bounded support. This might be complicated, but luckily we can generate samples from the 1-dimensional uniform distribution.

Let $A = \{(x, y) : 0 < y < p(x)\}$ the area below the curve of $p(x)$. Let $Z = (Z_x, Z_y) \in \mathbb{R}^2$ be a random variable with uniform distribution over $A \subset \mathbb{R}^2$.

1. Prove the the marginal distribution of Z_x is $p(x)$.

   Sampling Algorithm: To get samples from $p(x)$ let $x^{(0)} \in supp(p)$ arbitrary, and do the following iterations for $i = 0, 1, 2, \ldots$:

   - Sample $y^{(i+1)} \sim Uniform([0, p(x^{(i)})])$
   - Sample $x^{(i+1)} \sim Uniform(\{x : p(x) > y^{(i+1)}\})$

2. Draw a picture that illustrates how this sampling algorithm works. (Hint: The algorithm is called slice sampling)

3. Prove that the limit distribution of $x^{(1)}, x^{(2)}, \ldots$ is the distribution having density $p(x)$.

Solution:

1. First calculate the density of a uniform distribution over the domain A. Since the distribution is uniform, therefore the density is constant on A, for example $c > 0$. We also know that the integral of the density over A is 1. Therefore,

$$\int_{-\infty}^{\infty} \int_{0}^{p(x)} c \, dy \, dx = 1,$$

and

$$c = \frac{1}{\int_{-\infty}^{\infty} \int_{0}^{p(x)} \, dy \, dx} = \frac{1}{\int_{-\infty}^{\infty} p(x) \, dx} = \frac{1}{1} = 1.$$

The density of the random variable Z_x at point x is

$$\int_{0}^{p(x)} c \, dx = \int_{0}^{p(x)} \, dx = p(x)$$

From this it follows that if we want to draw samples from density $p(x)$, then it is enough to generate 2-dimensional uniformly distributed samples over A (i.e. $(Z_x, Z_y) \sim Uniform_A$), and then keep Z_x only.

2. See e.g. http://en.wikipedia.org/wiki/Slice_sampling

3. To see that the limit distribution of slice sampling is $p(x)$, it is enough to show that the sampling algorithm is a special case of Gibbs sampling where the limit distribution is uniform over A:

$$p(x, y) = \begin{cases} 1 & \text{if } (x, y) \in A \\ 0 & \text{otherwise.} \end{cases}$$

We have to show that

$$\begin{aligned} p(x|y) &= Uniform(\{x : p(x) > y\}) \\ p(y|x) &= Uniform([0, p(x)]) \end{aligned}$$

These can be easily seen from the following equations:

$$\text{If } x \in A \text{, then } p(y|x) = \frac{p(y, x)}{p(x)} = \frac{p(y, x)}{\int_0^{p(x)} p(y, x) \, dy} = \frac{1}{\int_0^{p(x)} 1 \, dy} = \frac{1}{p(x)}.$$

$$p(x|y) = \frac{p(y, x)}{p(y)} = \frac{p(y, x)}{\int_{-\infty}^{\infty} p(y, x) \, dx} = \frac{1_{\{(x,y) \in A\}}}{\int_{-\infty}^{\infty} 1_{\{(x,y) \in A\}} \, dx} = \frac{1_{\{x : p(x) > y\}}}{\int_{-\infty}^{\infty} 1_{\{x : p(x) > y\}} \, dx}.$$

10 Entropy, Mutual Information

Let $p_1, \ldots, p_d$ be a discrete distribution ($p_i \geq 0$, $\sum_{i=1}^d p_i = 1$). Let f be a density function of a continuous distribution.

1. Prove that $H(p) = -\sum_{i=1}^d p_i \log p_i \geq 0$, that is, the discrete entropy is nonnegative.
2. Give an example where $H(f) = -\int f \log f < 0$, that is the entropy of continuous distributions can be negative.

   Answers. 1. Note that $p_i \in [0, 1]$, and therefore $\log p_i \leq 0$, therefore $H(p) \geq 0$.

   2. There are several examples, for example, the entropy of a Gaussian distribution is $\frac{1}{2}\ln(2\pi e\sigma^2)$. Choose σ such that $\sigma^2 < 1/(2\pi e)$ we obtain negative entropy.

The conditional mutual information is defined as $I(X,Y|Z) = \mathbb{E}_Z(I(X,Y)|Z=z)$, where I denotes the mutual information.

3. Explain what the conditional mutual information means.
4. Prove that $I(X,Y|Z) = H(X,Z) + H(Y,Z) - H(X,Y,Z) - H(Z)$, where H denotes the entropy.
5. Give examples such that (a) $I(X,Y|Z) > I(X,Y)$, and (b) $I(X,Y|Z) < I(X,Y)$.

Answers. 3. the expected value of the mutual information of two random variables given the value of a third.

4. $I(X,Y|Z) = H(X|Z) - H(X|Y,Z) = H(X,Z) - H(Z) - (H(X,Y,Z) - H(Y,Z))$

5. (a) Consider $X \perp Y$, $Z = X \text{ xor } Y$. Then $I(X,Y) = 0$ where $I(X,Y|Z) > 0$ because given Z, we know X if Y is given.

(b) is the common case, for example $X \not\perp Y$, $Z \not\perp (Y,X)$. The proof can be see from a van diagram.

Another examples is the following structure:

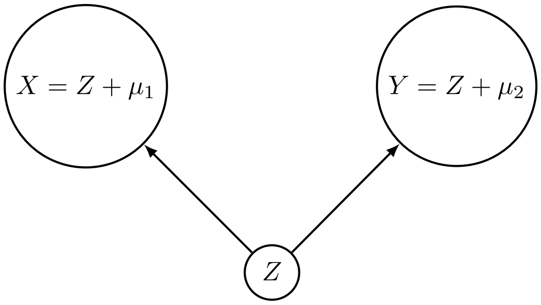


Figure 5: Common cause structure

11 Convexity

1. **Convex sets.** The convex hull of a set X is defined as:

$$\text{Conv}(X) = \{\sum_i a_i x_i \mid \forall i, a_i \geq 0 \;\&\; \sum_i a_i = 1\}$$

   - Show that no matter X is convex or not, Conv(X) is always a convex set.
   - Show that Conv(Conv((X)) = Conv(X)

2. **Convex functions.** In Jensen's inequality, if X is a random variable and ϕ is a convex function, then $\phi[EX] \leq E[\phi(X)]$. Prove this inequality when X is discrete and finite, i.e., X can take values from $\{x_1, x_2, ..., x_n\}$ with probability $\{p_1, ..., p_n\}$.

Solution: 1. (1) For any $X_1, X_2 \in$ Conv(X), and for any $0 \leq \lambda \leq 1$,

$$\lambda X_1 + (1-\lambda)X_2 = \lambda \sum_i a_i x_i + (1-\lambda)\sum_i b_i x_i = \sum_i (\lambda a_i + (1-\lambda)b_i)x_i$$

since $a_i, b_i \geq 0$ and $0 \leq \lambda \leq 1$, we know $(\lambda a_i + (1-\lambda)b_i) \geq 0, \forall i$, and since $\sum_i a_i = 1, \sum_i b_i = 1$,

$$\sum_i (\lambda a_i + (1-\lambda)b_i) = \lambda \sum_i a_i + (1-\lambda)\sum_i b_i = 1$$

Hence $\lambda X_1 + (1-\lambda)X_2 \in$ Conv(X). We can conclude that Conv(X) is always a convex set.

(2) For any element $y \in$ Conv(Conv(X)), suppose $\sum_j b_j = 1$ and $\sum_i a_{ji} = 1, \forall j$, then

$$y = \sum_j b_j (\sum_i a_{ji} x_i) = \sum_i \sum_j b_j a_{ji} x_i$$

The sum of coefficients is:

$$\sum_i \sum_j b_j a_{ji} = \sum_j b_j \sum_i a_{ji} = 1$$

and it is easy to show that the coefficients are nonnegative since both b_j and a_{ji} are nonnegative. Hence $y \in$ Conv(X), and therefore Conv(Conv(X)) $\subseteq$ Conv(X). On the other hand, it is easy to see that $X \subseteq$ Conv(X) since we can always take $a_i = 1, a_j = 0, j \neq i$ to pick out $x_i, \forall i$. Substitute X with Conv(X) we have Conv$(X) \subseteq$ Conv(Conv(X)). So we conclude Conv(Conv((X)) = Conv(X).

2. Use induction. The base case ($n = 2$) is easy to prove using the property of convex functions. Suppose the inequality holds when $n = k$, i.e.

$$\phi[p_1 x_1 + ... + p_k x_k] \leq p_1 \phi(x_1) + ... + p_k \phi(x_k)$$

where $p_1 + ... + p_k = 1$.
Now for $n = k + 1$, if $p_1 + ... + p_k + p_{k+1} = 1$, using the property of convex functions,

$$\begin{aligned}
\phi[p_1 x_1 + ... + p_k x_k + p_{k+1} x_{k+1}] &= \phi[(1-p_{k+1})\frac{p_1 x_1 + ... + p_k x_k}{1-p_{k+1}} + p_{k+1}x_{k+1}] \\
&\leq (1-p_{k+1})\phi(\frac{p_1 x_1 + ... + p_k x_k}{1-p_{k+1}}) + p_{k+1}\phi(x_{k+1}) \\
&\leq (1-p_{k+1})\frac{p_1 \phi(x_1) + ... + p_k \phi(x_k)}{1-p_{k+1}} + p_{k+1}\phi(x_{k+1}) \\
&= p_1 \phi(x_1) + ... + p_k \phi(x_k) + p_{k+1}\phi(x_{k+1})
\end{aligned}$$

Hence we conclude that $\phi[EX] \leq E[\phi(X)]$.

12 Exponential Families

Let $\mathcal{P}$ be a set of distributions that contains $U[0,1]$, the uniform distribution on $[0,1]$.

- Give an example when $\mathcal{P}$ contains infinitely many distributions and $\mathcal{P}$ is an exponential family.
  Answer The set of beta distributions.
- Give an example when $\mathcal{P}$ contains infinitely many distributions and $\mathcal{P}$ is NOT an exponential family.
  Answer The set of uniform distributions $U[1,\theta]$.